(1)

Technical Report 905

Human and Computer Task Allocation in Air Defense Systems: Final Report

Marvin S. Cohen, Leonard Adelman, Terry A. Bresnick,
James O. Chinnis, Jr., and Kathryn B. Laskey
Decision Sciences Consortium, Inc.

August 1990

FIFTY YEARS OF SERVICE
1940 - 1990

United States Army Research Institute
for the Behavioral and Social Sciences

U.S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON **Technical Director**	**JON W. BLADES** **COL, IN** **Commanding**

Research accomplished under contract
for the Department of the Army

Decision Sciences Consortium, Inc.

Technical review by

Laurel Allender
Randall M. Chambers

NOTICES

~~DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POX, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.~~

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION: Unclassified	1b. RESTRICTIVE MARKINGS: --
2a. SECURITY CLASSIFICATION AUTHORITY: --	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE: --	
4. PERFORMING ORGANIZATION REPORT NUMBER(S): 87-16	5. MONITORING ORGANIZATION REPORT NUMBER(S): ARI Technical Report 905
6a. NAME OF PERFORMING ORGANIZATION: Decision Science Consortium, Inc. — 6b. OFFICE SYMBOL (If applicable): --	7a. NAME OF MONITORING ORGANIZATION: U.S. Army Research Institute for the Behavioral and Social Sciences
6c. ADDRESS (City, State, and ZIP Code): 1895 Preston White Drive, Suite 300, Reston, VA 22091	7b. ADDRESS (City, State, and ZIP Code): 5001 Eisenhower Avenue, Alexandria, VA 22333-5600
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: U.S. Army Research Institute for the Behavioral and Social Sciences — 8b. OFFICE SYMBOL (If applicable): PERI-SM	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: MDA903-85-C-0332
8c. ADDRESS (City, State, and ZIP Code): 5001 Eisenhower Avenue, Alexandria, VA 22333-5600	10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
665502M	770	N/A	N/A

11. TITLE (Include Security Classification)
Human and Computer Task Allocation in Air Defense Systems: Final Report

12. PERSONAL AUTHOR(S) Cohen, Marvin S.; Adelman, Leonard; Bresnick, Terry A.; Chinnis, James O., Jr.; and Laskey, Kathryn B.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM 85/09 TO 87/10	1990, August	149

16. SUPPLEMENTARY NOTATION
Contracting Officer's Technical Representative (COTR), Irving Alderman; alternate COTR, Laurel Allender.

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Decision aids; Man-computer systems; Man-machine interface; Decision making

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The first task of the Phase II research involved additional analysis of the Phase I experimental data with the goal of developing models of operators' information-processing strategies when using different human-computer interfaces, and the relationship between those models and performance. The second task involved: (1) the development of a representative, computer-based testbed for performing controlled, experimental research with actual U.S. Army air-defense operators; and (2) the performance of two experiments at Fort Bliss for testing (a) the relative effectiveness of alternative interfaces for supporting human-computer interaction, (b) the theoretical principles underlying the predictions, and (c) the ability to link information-processing strategies to performance.

The research demonstrated the superior performance of interfaces that solved the relatively easier tasks and helped operators focus their attention on the relatively harder tasks under conditions of high workload. In addition, the research demonstrated the added

(Continued)

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION: Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL: Irving Alderman	22b. TELEPHONE (Include Area Code): (202) 274-9134 — 22c. OFFICE SYMBOL: PERI-SM

ARI Technical Report 905

19. ABSTRACT (Continued)

value achieved by an operator-controlled allocation (i.e., rule creation) capability that permits the operator to instruct the system in how to perform certain tasks (i.e., identify targets). A direct relationship between operators' performance and their information-processing strategies was demonstrated for each interface.

Keywords: decision aids, decision making. (KR)

Accession For	
NTIS GRA&I	☑
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

QUALITY INSPECTED 1

Technical Report 905

Human and Computer Task Allocation in Air Defense Systems: Final Report

Marvin S. Cohen, Leonard Adelman, Terry A. Bresnick,
James O. Chinnis, Jr., and Kathryn B. Laskey
Decision Sciences Consortium, Inc.

Manned Systems Group
John F. Hayes, Chief

Systems Research Laboratory
Robin L. Keesee, Director

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

August 1990

Army Project Number
2P665502M770

Small Business Innovative
Research

FOREWORD

The Manned Systems Group of the Army Research Institute (ARI) conducts research and development in areas concerned with manpower, personnel, and training issues in systems development. A critical issue is to develop technology for the better allocation of tasks between soldier and computer in systems requiring complex decision making. Anticipated high levels of performance have not been achieved in recent systems development efforts.

The research reported herein investigates experimentally the performance and effectiveness of several alternative allocation schemes under controlled conditions of workload and uncertainty. The research paradigm is applicable to command and control systems in general. It provides an initial knowledge base for better allocation of tasks and improved design of the soldier-computer interface in future systems.

This research was conducted under the Small Business Innovative Research (SBIR) program for research and development on DoD scientific and engineering problems. It is based on an earlier demonstration of feasibility and is a potential precursor to transfer of the technology to the civilian sector.

EDGAR M. JOHNSON
Technical Director

ACKNOWLEDGMENTS

This work has been supported by the U.S. Army Research Institute for the Behavioral and Social Sciences, under the Defense Small Business Innovative Research (SBIR) program. We are grateful to our scientific officers, Irving Alderman and Laurel Allender, for helpful guidance and assistance; to Russell J. Branaghan and Mark J. Sugg, graduate students in the Department of Psychology at New Mexico State University, who were the experimenters for the two experiments at Fort Bliss; to our colleagues at Decision Science Consortium, Inc. (DSC), John Leddo, Jeffrey S. Mandel, F. Freeman Marvin, Martin A. Tolcott, for significant contributions to the work, Lynn Merchant-Geudero, for handling contract-management aspects of the effort, and particularly to Monica M. Govedich and Bryan B. Thompson, the two DSC computer scientists who developed the air defense testbed for experimentation.

HUMAN AND COMPUTER TASK ALLOCATION IN AIR DEFENSE SYSTEMS: FINAL REPORT

EXECUTIVE SUMMARY

Requirement:

The advent of increasingly sophisticated and expensive human-machine systems has called into question basic assumptions about the proper respective roles of computers and humans. In particular, the reallocation of cognitive tasks from human to computer has sometimes resulted in user rejection of resulting systems or in systems that may not take full advantage of human contributions to the overall task. Research is, therefore, required to: (1) directly test hypotheses about human-computer interaction with the goal of determining the relative effectiveness of alternative types and levels of interface capabilities for allocating tasks between the human and computer; (2) develop a more fundamental theoretical understanding of the psychological mechanisms underlying human-computer performance; and (3) move toward the development of a "cognitive" human factors technology for predicting human-computer system performance on the basis of information-processing models.

Procedure:

The first task of the Phase II research involved additional analysis of the Phase I experimental data with the goal of developing models of operators' information-processing strategies when using different human-computer interfaces, and the relationship between those models and performance. The second task involved: (1) the development of a representative, computer-based test-bed for performing controlled, experimental research with actual U.S. Army air defense operators; and (2) the performance of two experiments at Fort Bliss for testing (a) the relative effectiveness of alternative interfaces for supporting human-computer interaction, (b) the theoretical principles underlying the predictions regarding the effectiveness of the interfaces, and (c) our ability to link information-processing strategies to performance.

Findings:

The research demonstrated the superior performance of interfaces that solved the relatively earlier tasks and helped operators focus their attention on the relatively harder tasks under conditions of high workload. In addition, the research demonstrated the clear, added value achieved by an operator-controlled allocation (i.e., rule creation) capability that permits the operator to instruct the system in performing certain tasks (i.e., target identification), thereby freeing the operator to gather more information and take longer to examine those targets requiring his/her attention. A direct relationship between operators' performance with each interface and their information-processing strategies was demonstrated in all cases.

Utilization of Findings:

The research has implications for the development of general guidelines for constructing human-computer interfaces for performing tasks involving the identification of a large number of objects (e.g., aircraft) under conditions of high workload and uncertainty. In addition, the findings and broader theoretical and methodological approach are applicable to Army domains other than air defense.

HUMAN AND COMPUTER TASK ALLOCATION IN AIR DEFENSE SYSTEMS: FINAL REPORT

CONTENTS

	Page
INTRODUCTION	1
ADDITIONAL ANALYSES OF PHASE I DATA	7
Phase I: Experimental Method and Principal Results	7
Additional Analyses and Cognitive Modeling	15
PHASE II EXPERIMENTS WITH ACTIVE-DUTY AIR DEFENSE OPERATORS	35
Basic Principles and a General Experimental Design	35
The Air Defense Task and Experimental Testbed	39
Shafer's Theory of Evidence	47
Experiment 1	50
Experiment 2	69
CONCLUSIONS	95
Toward Guidelines for Human-Computer Interfaces	96
Implications for Army Command and Control	97
REFERENCES	105
APPENDIX A. INSTRUCTIONS DESCRIBING THE PARTICIPANTS' "BASIC JOB" AND THE AIR DEFENSE TESTBED	107
B. THE "SYSTEM CAPABILITY" DESCRIPTIONS FOR THE MANUAL, OVERRIDE, AND SCREENING INTERFACES IN EXPERIMENT 1	115
C. THE SUBJECTIVE QUESTIONNAIRES USED IN BOTH EXPERIMENTS	119
D. THE "SYSTEM CAPABILITY" DESCRIPTIONS FOR THE THREE ALLOCATION CONDITIONS IN EXPERIMENT 2	125
E. CUE-PROCESSING SEQUENCE FOLLOWED BY ONE OF THE DSC DOMAIN EXPERTS IN THE NONALLOCATION CONDITIONS FOR THE MANUAL, OVERRIDE, AND SCREENING CONDITIONS	139

Page

LIST OF TABLES

Table 2-1. Diagnosticity of cue patterns 10

2-2. Diagnosticity of cue patterns for the four computer cues 11

2-3. Optimal responses and percent correct IDs for each cue pattern. 12

2-4. Percent optimal system responses, mean of all participants. 13

2-5. Overall analysis of variance. 13

2-6. Subsidiary analysis of variance 14

2-7. Equivalent conditions (shaded). 14

2-8. Probability a majority of sampled cues are hostile. . . . 23

2-9. Percent engaging of target by cue type. 27

2-10. Computer symbols and required operator responses to change symbols in the different interface conditions. 28

2-11. Response latency, manual condition. 32

2-12. Response latency, screening condition 32

2-13. Response latency, override condition. 33

3-1. Design for the first Phase II experiment. 36

3-2. Design for the second Phase II experiment 37

3-3. Hypothesized rank order of performance (in high workload) for different conditions 38

3-4. Degree of belief values for individual cue levels (assuming cue independence). 54

3-5. The IFF challenge and HQ ID responses constructed for one type of target. 55

Page

Table 3-6. Diagnosticity values for individual cue levels (based on all targets in performance simulation). 56

3-7. Experiment 1: The order of experimental conditions used for counterbalancing. 60

3-8. Experiment 1: The "mean percent correct" for N=58 targets with initial degree of belief below .60 62

3-9. Experiment 1: The "mean percent correct" for N=57 targets with initial degree of belief between .6 and .8. 63

3-10. Experiment 1: The "mean percent correct" for N=115 targets with initial degree of belief below .80 63

3-11. Experiment 1: The "mean percent correct" for "message targets" 64

3-12. Experiment 1: The "mean percent correct" for N=85 targets with initial degree of belief ≥ .80. 64

3-13. Experiment 1: The "mean percent correct" for the entire 200-target simulation. 64

3-14. Percent correct for a totally automated system (i.e., with initial data only) for different classes of targets. 65

3-15. Experiment 1: Mean subjective performance, workload, and preference (first questionnaire). 66

3-16. Mean utility rating for the target identification cues. 68

3-17. Rule work sheet 73

3-18. Experiment 2: The order of experimental conditions used for counterbalancing 75

3-19. Experiment 2: The "mean percent correct" for N=58 targets with initial degree of belief ≤ .60 77

3-20. Experiment 2: The "mean percent correct" for N=57 targets with initial degree of belief between .6 and .8 78

		Page
Table 3-21.	Experiment 2: The "mean percent correct" for N=115 targets with initial degree of belief below .80	79
3-22.	Experiment 2: The "mean percent correct" for N=56 "message targets"	80
3-23.	Experiment 2: The "mean percent correct" for N=85 targets with initial degree of belief $\geq$ .80	81
3-24.	Experiment 2: The "mean percent correct" for the entire 200-target simulation	81
3-25.	Mean percentage of targets hooked by class	83
3-26.	Mean length of time (in seconds) of targets hooked by class	85
3-27.	Experiment 2: Percentage of time different cues were requested by condition	87
3-28.	The types of rules created by three (or more) participants when using the allocation capability	88
3-29.	Experiment 2: Mean values for both objective workload measures for the five interfaces	89
3-30.	Experiment 2: Mean subjective performance, workload, and preference (first questionnaire)	90
3-31.	Experiment 2: Mean subjective performance, workload, and preference (second questionnaire; higher values represent better scores)	91
3-32.	Experiment 2: Mean utility rating for the target identification cues	92
D-1.	Rule work sheet	127
D-2.	Rule work sheet	132
D-3.	Rule work sheet	136

Page

LIST OF FIGURES

Figure 2-1.	Diagram of simulation process.	9
2-2a.	Optimal processing strategy for manual condition.	16
2-2b.	Majority of confirming dimensions strategy for manual condition	17
2-3a.	"Optimal" processing strategy for manual condition.	18
2-3b.	"Majority of confirming dimensions" strategy for manual condition	19
2-4.	Percentage of targets labeled hostile as function of number of hostile cues (manual, low workload).	22
2-5.	Performance in manual condition for low and high workload.	24
2-6.	Regression coefficients against cue values	26
2-7.	Mean response latencies for participants in the manual condition as a function of the number of hostile cues	34
3-1.	Hypothesized performance levels in high workload for the different conditions	39
3-2.	The radar screen and basic symbols	43
3-3.	The entire display screen.	45
3-4.	Initial screen display	72
3-5.	Mean percentage of targets hooked by interface condition and initial degree of belief	84
3-6.	Mean length of time (in seconds) of targets hooked by interface condition and initial degree of belief	86

		Page
Figure A-1.	The radar screen and basic symbols	108
A-2.	The entire display screen	110
D-1.	Initial screen display	126
D-2.	Initial screen display	131
D-3.	Initial screen display	135

HUMAN AND COMPUTER TASK ALLOCATION IN AIR DEFENSE SYSTEMS: FINAL REPORT

1.0 INTRODUCTION

There has been a tendency in the design of increasingly sophisticated human-computer systems to assign all possible tasks to the computer, leaving the human operator with a smaller and smaller role. There are many reasons for this trend, but a major one may be our lack of knowledge about how to design a system that makes maximum simultaneous use of both the human and the computer. The empirical findings reported in Phase I of the current program of research (Chinnis, Cohen, and Bresnick, 1984) demonstrated how a cognitive psychology focus could generate the knowledge needed to identify task-allocation design principles for complex human-computer systems. Although the experimental representation of the Army air defense system employed was much simpler than any real system, the results suggest that some strategies for task allocation will work better than others and that, in some settings under high processing loads, neither fully manual nor fully automated systems work as well as a scheme permitting some collaboration between the system's human and computer elements.

Under low-workload conditions, and where the computer model in the target identification (ID) task was significantly incomplete, any of the three conditions involving human participation was superior to computer-only performance in the Phase I research. However, at high workload, this advantage could be maintained only if the computer assisted the human in the function of "allocating tasks," i.e., by directing the user's attention to subproblems where his/her contribution was most needed. In short, a complementarity had been observed; of the four very different task-allocation systems tested, the only one that enabled both the computer and the human operator to make maximum use of their respective contributions was a "screening" condition that utilized the high-speed capability of the computer to process all aircraft according to its available (and programmed) information resources and then to signal the operator to attend to those aircraft it was unable to classify reliably.

While the importance of the role of a particular human operator in a complex and highly automated distributed or hierarchical system is not always easy to determine, in the case of more localized systems or in the case of breakdowns in the networking of distributed systems, the ability of the human operator to supply information to the system and affect system behavior may become highly desirable or even essential. The Phase I research had successfully attempted to shed some light on how that capability for human-machine interaction might best be achieved and to suggest the general direction of further research. The work performed in the first year of Phase II was directed toward laying the groundwork for extending the Phase I findings through a series of theory-driven experiments (in Year 2) using U.S. Army personnel performing representative air defense exercises and using a computer-based testbed that permits researchers to vary human-machine allocation schemes.

At the broader level, the objectives of Phase II were to carry the Phase I work forward along two fronts:

(1) primarily, toward developing a design technology for complex human-computer systems; and

(2) secondarily, toward deriving specific recommendations that might be incorporated into the design and modification of future Army air defense systems, as well as other Army domains.

To the extent possible, the design technology to be developed would consist not merely of a set of qualitative (though possibly useful) guidelines, but rather, it would be distinctive in three respects: (a) it would be based firmly on findings and theories in cognitive psychology; (b) it would be tested and confirmed by experimental evidence; and (c) it would include, to as great an extent as possible, procedures and formulae for quantitatively predicting human-computer performance under a variety of conditions. In short, the long-term goal is a technology that permits designers to link information-processing models to predicted performance that is applicable across a wide range of human-computer systems, as well as a set of specific proposed improvements in the performance of Army air defense systems based on that technology.

We argued in our Phase I work that traditional methods of task allocation are neither fine-grained nor flexible enough for many important applications (as noted, for example, by Singleton, 1974; Rouse, 1977; Cohen et al., 1982). The development of an alternative approach, however, is itself not without difficulties. On a more specific level, therefore, Phase II research addressed a variety of theoretical and methodological challenges that represent initial efforts to achieving a fully adequate technology for cognitive task allocation. In particular, these challenges were addressed to the following broader questions:

(1) *What is the impact of workload on human processing of information relevant to a decision?* Performance may suffer under low workload, since the human's ability to participate where required may be compromised by decreased vigilance. At the other extreme, however, as the number or complexity of decisions requiring human input increases, human performance is also likely to degrade. In particular, there is evidence that humans shift from reliance on more nearly optimal decision rules under low workload to use of suboptimal simplifying strategies or heuristics under high workload. For example, as workload increases, decisions may be made by comparing options to cut-off points on one or a small number of relevant dimensions, rather than evaluating each option with respect to all available cues (Payne, 1976). Under these conditions, the relative advantage of humans over computers, even when the computer's model is incomplete, may be lost.

(2) *What is the impact of allocation schemes that represent cooperative problem solvers?* Although the Phase I results were regarded as tentative, they suggested that provision of a capability for the operator to override computer decisions was suboptimal when compared with methods which were more collaborative in nature. Of the four very different task-

allocation systems tested, the only one which enabled both the computer and the human operator to make maximum use of their respective contributions was one which utilized the high-speed capability of the computer to process all aircraft according to its available (and programmed) information resources and then to signal the operator to attend to those aircraft it was unable to classify reliably.

(3) *How do flexible allocation schemes compare with fixed ones?* Fixed schemes involve pre-set allocation of tasks to the human and the computer, while flexible schemes may vary the computer and human contributions in any decision, depending on situational variables such as workload and the relative expertise of the user and computer. For example, in a relatively low-workload situation, an attack planning or targeting C^2 system might allocate data collection and display functions to the computer and leave inferences and predictions regarding critical events (e.g., identity and intentions of hostile and friendly contacts), target selection, and decisions to engage or not to engage to the human operator. Under a high-workload multi-threat situation, however, the system "executive" might reallocate more of the integrative tasks (Phelps, Halpin, and Johnson, 1981) to the computer. As stress and load increase, for example, the computer might begin to display recommended attack plans, target priorities, and weapon-target assignments. Under still higher stress and load, the computer might assume control of the actual firing of weapons or configuration of combat equipment. We were particularly interested in the effects of giving operators the flexibility to allocate tasks to the machine on-line by creating "rules" for governing the inference process implemented by the machine.

Experimental work addressing these questions was conducted in a framework similar to that developed in the Phase I research, but with a more representative air defense testbed and with actual air defenders. We distinguish, however, two components of our approach. The first involves the direct testing of hypotheses about human-computer interaction. For example, one such hypothesis is that a dynamic screening condition (such as the one tested in Phase I) is appropriate under conditions of high workload and where the human may have knowledge not possessed by the computer. The second component, however, is the development of a more fundamental theoretical understanding of the psychological mechanisms underlying human-computer performance. It is this second thrust which makes possible a more integrative, more generalizable, and more quantitative design technology than the mere accumulation of isolated principles or guidelines.

The central idea is that, by understanding and modeling operators' information-processing strategies, we will be able to develop technology in which overall human-computer system performance could be *predicted*, under diverse task-allocation conditions and under diverse conditions of workload, threat, cue conflict, and other variables. Such predictions will often, of course, be approximate and relative; they will not remove the need for exploratory implementation, experimentation, and testing of new

systems. Nonetheless, a theory-based, quantitative methodology such as this could represent a quantum leap in the ability of engineers and system designers to make use of psychological results in the design of human-computer systems.

Given the above perspective, the first task performed in the first year of Phase II was additional analysis of the Phase I data and modeling of those data in terms of information-processing strategies. This analysis focused on a more extensive analysis of the basic performance variable (target ID decisions), regression analysis of participants' (Ps') reliance on different subsets of cues, and response latencies. The procedures and results obtained for these analyses are presented in Section 2.0 of this report.

The second major task, which comprised the bulk of the Phase II effort, was the development and implementation of a computer-based testbed that permitted controlled experimentation regarding the relative effectiveness of alternative human-machine allocation schemes with actual U.S. Army personnel performing representative air defense exercises. This broad task had a number of primary subtasks. For example, the research team had to identify alternative human-machine allocation schemes that were both theoretically distinct and interesting, and yet operationally definable within the context of our testbed. Second, we had to learn enough about the Army air defense problem so that we could: (a) design a representative testbed; (b) assess the implications of trading off certain characteristics of the "real air defense problem" (e.g., when and how information is presented) in order to obtain the experimental control necessary to evaluate Ps' information-processing strategies in alternative conditions; and (c) incorporate substantive domain knowledge into the testbed so that the machine appropriately processed the information available to it. A third subtask was to extend the machine's capability for processing information well beyond that in Phase I so that it could deal with representative real-world conditions regarding lack of information and conflicting information. Toward this end, Shafer's (1976) theory of evidence was used as a pragmatically justifiable basis for expanding the "screening" condition in Phase I into a more collaborative, dialogue- (or suggestion-) oriented human-machine allocation scheme for investigation in Phase II. A fourth subtask in developing the computer-based testbed was the design and development of the simulated air defense scenarios for the participants to execute in the experiment. A fifth subtask was the programming necessary to actually implement the testbed on an IBM PC-AT. The sixth subtask involved all the activities required for successfully conducting two experiments using the testbed with actual U.S. Army air defense operators at Fort Bliss, Texas. The first experiment was basically a replication effort of the Phase I experiment, but now with a more representative task and participants. The second experiment extended the first experiment by evaluating flexible allocation capabilities implemented by on-line rule creation. The eighth subtask involved performing the detailed and lengthy analysis of the data collected at Fort Bliss. Analysis of the second experiment focused, in particular, on the relationship between performance and information-processing strategies for different human-machine interfaces. All of these subtasks are discussed in Section 3.0 of the report.

Section 4.0 of the report discusses the results of the experiments from the perspective of (a) developing guidelines for human-machine interfaces, and (b) extending both the findings and broader theoretical and methodological approach to other Army domains.

2.0 ADDITIONAL ANALYSES OF PHASE I DATA

The first task performed in the first year of Phase II was a more detailed analysis of the Phase I data and preliminary modeling of the data obtained for the different human-computer allocation conditions in terms of information-processing strategies. This analysis focused on a more refined analysis of ID performance data, the extent to which participants relied on different subsets of the five available cues, and their response latencies in the different experimental conditions. Prior to presenting the procedures and results for these additional analyses, the Phase I experimental method and principal results are reviewed below. A complete description of the Phase I experimental materials and results can be found in Chinnis, Cohen, and Bresnick (1985).

2.1 Phase I: Experimental Method and Principal Results

2.1.1 Experimental method. The task for the combined participant-computer system was to observe approaching aircraft and various data regarding the aircraft and to make decisions whether to shoot or not shoot each aircraft based on whether the aircraft was friendly or hostile. Two computer-driven displays were used to simulate an air defense console and to present information to the participants (Ps).

Aircraft symbols appeared at the top of one display (the "radar" display) moving at constant and identical speed toward the bottom center of the display, where the P and the air defense system were "located." Traversal of the screen from the top of the display to an "in-range" line, where missiles were fired, required approximately one minute. Four cues were available to *both* the computer and P. One of these was available from the graphics ("radar") screen, the other three from the companion text screen. In addition, a fifth cue--referred to as the "extra" cue--was available *only* to P; it was available from the graphics screen, but was not utilized in the computer's ID algorithm. Information provided in the instructions (regarding the overall location of friendly and hostile air bases) was intended to alert Ps to the potential significance of this cue.

The experiment used a within-Ps design with two primary independent variables or treatments. The first is the task allocation, consisting of three conditions:

- A *manual* condition, in which all aircraft are initially shown as having "unknown" ID on the display screen and in which Ps must make all ID decisions. This is a human-only allocation scheme.

- An *override* condition, in which the computer applies a Bayesian algorithm utilizing four cues to determine target ID and labels all targets as either hostile or friendly, but in which Ps may override the computer in those cases which they believe are wrong.

- A *screening* condition, in which the computer processes the four cues available to it for each aircraft, but produces an ID decision only when 3 out of 4 of these cues agree. The computer requests P assistance (by means of an "unknown" ("?") symbol) for cases of cue conflict (i.e., where the Bayesian ID algorithm in the computer was inconclusive).

A fourth condition, useful for comparison purposes, is the fixed, computer-alone scheme. In this condition, system performance measures can be calculated analytically, without resort to experimental data.

The second independent variable is information load. This was used at two levels. Low load resulted in six aircraft simultaneously on the screen (a new aircraft appearing every 11 seconds), whereas high load resulted in fifteen simultaneous aircraft (a new aircraft appearing every 4 seconds). These aircraft counts refer to aircraft positioned above the "in-range" line, i.e., those far enough away from the P that decisions (shoot or don't shoot) can still be made. A few aircraft are usually on the screen beyond this point.

In all cases the P's mode of response was to type the number associated (on the display screen) with an aircraft. The result of this action varied among the different task-allocation conditions. In the manual condition, all aircraft were shown as unknowns; typing its target number caused an aircraft to be designated hostile and to be destroyed. In the override condition, all aircraft were identified by the computer as either friendly or hostile; typing a target's number caused this designation to reverse. As before, hostile aircraft were destroyed. In the screening condition, aircraft associated by the computer with "unknown" symbols were treated as in the manual condition: i.e., targets whose numbers were typed were designated hostile and destroyed. Other aircraft, identified by the computer as either friendly or hostile, were treated as in the override condition: i.e., typing its number reversed the designation. Whenever an aircraft number was entered, the aircraft symbol shown on the radar screen was modified by enclosing it in a hexagon.

Feedback was provided to the Ps in three simultaneous ways:

- by the use of a flashing aircraft symbol in the lower part of the radar screen to indicate an error (friendly destroyed, or hostile not destroyed);

- by the use of a "right" or "wrong" message displayed next to the appropriate aircraft data line on the text screen;

- by a running score displayed on the text screen in the form of "number of correct decisions out of number attempted."

Ps were not in the military. Ps were paid six cents per aircraft correctly classified, leading to average earnings of approximately $9.00 per hour.

A counter-balanced approach was taken to the two independent variables in the within-Ps design. All Ps participated in three two-hour ses-

sions, with each session devoted to one of the task allocation conditions. In each session, each P participated first in a training session followed by a high-load condition and low-load condition in either order. Twenty-four Ps participated. Ps were recruited from the local area using bulletin board notices and other means.

For each P, and each of the six combinations of task-allocation condition and load condition, 200 responses (aircraft classifications) were obtained and used as data. The first and last 25 responses out of each 250-response cell were discarded due to transient effects related to starting and terminating each data-collection segment. Thus, a total of 28,800 usable responses was obtained.

Simulation and Cue Diagnosticity. Emphasis was placed on presenting Ps with a discrimination problem that was as representative of the air defense environment as possible. Therefore, although primary interest resided in the handling of discriminations for which substantial conflict of ID cues was present, a simulation was developed which provided an abstract but comprehensive situation from which Ps could learn about cue diagnosticities. No direct information was provided to Ps regarding the relative usefulness of cues; instead, they were provided with training prior to the experimental trials intended to enable them to extract the required information for themselves. This training consisted of a complete practice block of 250 trials at the start of each session, with a load intermediate between the low- and high-load conditions.

The computer software utilized a simulation module which generated a sequence of aircraft and ID cues according to the probability diagram shown in Figure 2-1.

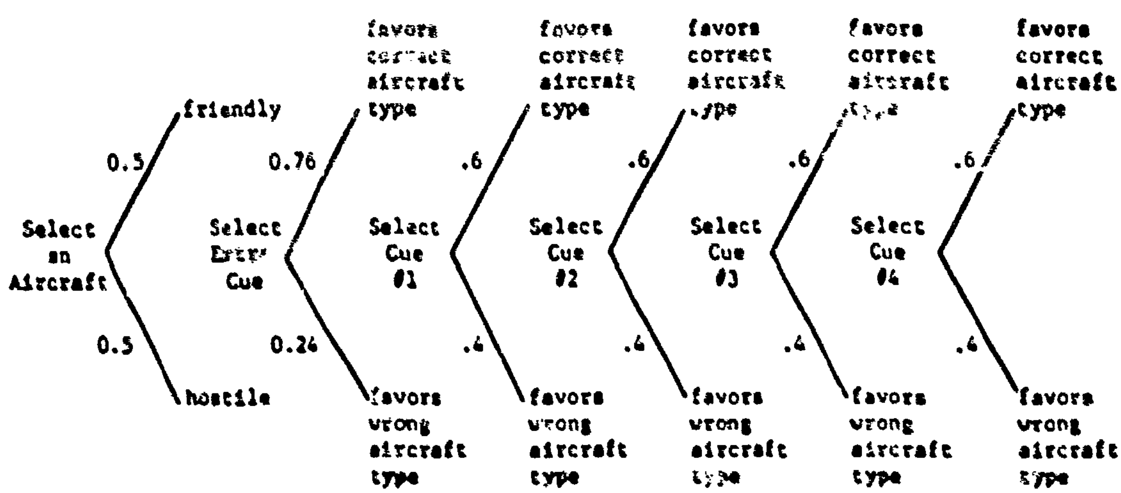


Figure 2-1. Diagram of simulation process.

The simulation generated friendly and hostile aircraft with equal probability. Following the selection of the true aircraft type, each of the four

cues utilized by the computer aid was independently generated so as to be appropriate to the type of aircraft with probability 0.6. Thus, for example, a friendly aircraft was assigned to fly within a safe-passage corridor with probability 0.6, whereas a hostile aircraft was assigned to a safe-passage corridor with probability 0.4. The "extra" cue--not utilized by the computer aid, but available to Ps who learned it--was generated similarly, but was somewhat more reliable in that it was selected to favor the true aircraft type with probability 0.76. This meant, for instance, that a hostile aircraft was positioned on the left side of the screen (where the hostile base was located) with probability 0.76.

The simulation process utilizing 5 binary cues resulted, of course, in 32 possible cue combinations for any aircraft. Since the four cues available to the computer aid were equivalent in diagnosticity, this is represented more simply as a combination of (1) the number of the four computer cues that favor hostile, and (2) the extra cue. This representation is used throughout this report. Schematically it is shown in Table 2-1:

Table 2-1

Diagnosticity of Cue Patterns

Number of Computer Cues Favoring Hostile	Extra Cue Favors	Probability of Cue Pattern if Aircraft is		Hostile/Friendly Likelihood Ratio of Cue Pattern
		Friend	Hostile	
0	f	0.0985	0.0061	0.062
1	f	0.2627	0.0369	0.140
2	f	0.2627	0.0829	0.316
0	h	0.0311	0.0195	0.627
3	f	0.1167	0.0829	0.710
1	h	0.0829	0.1167	1.408
4	f	0.0195	0.0311	1.595
2	h	0.0829	0.2627	3.169
3	h	0.0369	0.2627	7.126
4	h	0.0061	0.0985	16.147
		1.0	1.0	

The optimal strategy in this context depends, of course, on the scoring system. In this study, there was an equal penalty for either type of error--shooting down a friendly or failing to shoot a hostile. Therefore, the optimal decision rule was to respond "hostile" in all cue patterns where the likelihood ratio exceeded 1--i.e., for all patterns that were more likely under the assumption of a hostile aircraft than under the assumption of a friendly aircraft.

In the table, notice that the extra cue--available only to Ps (who must infer it)--is sufficiently diagnostic to overwhelm all of the other four cues except when they all agree. In other words, optimal performance by a P would correspond to always responding according to the extra cue, unless all four of the other cues favor the other aircraft classification.

A similar table (Table 2-2) has been calculated for the computer aid alone--i.e., for use of the four cues without the extra cue.

Table 2-2

Diagnosticity of Cue Patterns for the Four Computer Cues

Number of Computer Cues Favoring Hostile	Probability of Cue Pattern if Aircraft is Friend	Hostile	Hostile/Friendly Likelihood Ratio of Cue Pattern
0	.130	.023	.77
1	.346	.154	.445
2	.346	.346	1.000
3	.154	.346	2.247
4	.023	.130	5.652

The optimal computer performance based on these four cues alone is to choose the aircraft identification suggested by the majority of cues: if more than 2 cues favor hostile, shoot; if fewer than 2 cues favor hostile, do not shoot; if 2 cues favor hostile, either response is equally likely to be correct.

Use of 5 cues rather than 4 will be an advantage in this task when the four computer cues do not agree. Note, however, that since no cue or cue pattern is perfectly associated with hostile or friendly, ID "errors" would still be expected even by an optimally performing participant. Thus we need to distinguish between the theoretically appropriate or optimal response (the "best decision") and the response which happens to be correct on a given occasion (a "good outcome"). An optimal response rule will produce fewer mistaken IDs *on the average* (hence, a higher overall score) but will not be right every time.

Table 2-3 shows the optimal response for each pattern of cues and the percentage of correct IDs that would result. These figures are given both for the computer (which has access only to four cues) and for the total human-computer system (which has access to four cues plus the extra cue). Asterisks indicate conditions under which utilization of four cues plus the extra cue may lead to a different ID decision than use of four cues alone. Note also that while use of the extra cue will help in cases of 1 cue pointing one way and 3 cues the other, it helps more when 2 cues point each way (the case of maximum ID conflict for the automated system).

Table 2-3

Optimal Responses and Percent Correct IDs for Each Cue Pattern

Number of Computer Cues Favoring Hostile	Extra Cue Favors	4 Cues Only		4 Cues Plus Extra Cue	
		Optimal Response	% Correct IDs	Optimal Response	% Correct IDs
0	f	f	94.13	f	94.13
1	f	f	87.69	f	87.69
2	f	-	50.00	f*	76.00
0	h	f	61.52	f	61.52
3	f	h	41.54	f*	58.46
1	h	f	41.54	h*	58.46
4	f	h	61.52	h	61.52
2	h	-	50.00	h*	76.00
3	h	h	87.69	h	87.69
4	h	h	94.13	h	94.13

2.1.2 Principal results. Most of the trials in the experiment were necessary to produce an ecologically valid set of stimuli which could be regarded as reasonably representative of the air defense environment and from which participants could learn the diagnostic values of cues. The primary result of interest, however, has to do with the handling by the human-computer system of those cases involving a high level of ID cue conflict. Since we are interested to see which task-allocation conditions, if any, enable better integration and utilization of both computer and human contributions to the problem, attention will focus below on the case of two of the cues available to the computer pointing toward hostile and two pointing toward friend.

The dependent variable is the appropriateness of the human-computer system response to each aircraft--in this case, shooting when the extra cue favors hostile and not otherwise. The data available consist of six scores per participant: a score for both high and low workload in each of three task-allocation conditions (manual, screening, override). Each score is the percentage of appropriate responses out of 68 observations. Since there are 24 participants, we have almost 10,000 total observations.

Table 2-4 summarizes these data. It shows the percentage of responses that were appropriate in each of the six conditions, averaged across participants. For each cell in the table, the highest achievable score is 100%. (A score of 100% would mean that all responses had conformed to the optimal decision rule for the 2 vs. 2 cue conflict condition, i.e., to respond in the direction of the extra cue. It does not mean that all responses would, in fact, have been correct identifications.) Data for the computer conditions were derived analytically, and reflect the 50% optimal response rate expected when there is no knowledge of the extra cue. This rate is equivalent to chance performance. 50% reflects the level of performance expected by chance alone in all cells of the table.

Table 2-4

Percent Optimal System Responses,
Mean of all Participants

	MANUAL	SCREENING	OVERRIDE	COMPUTER
LOW LOAD	74.57	72.24	71.69	50.00
HIGH LOAD	61.34	70.89	64.46	50.00

Note that the percentage *correct* responses is a linear function of the percentage of *optimal* responses. Thus, although the analyses to be presented here are in terms of optimal responses, the results would be essentially unchanged in an analysis based on correct responses.

The most salient observation regarding these data is the superiority of *all* the conditions in which human participation occurred in comparison to the computer-only condition. T-tests were performed comparing the mean number of appropriate responses in each of the six experimental conditions with the 50% optimal response rate expected of the computer algorithm (i.e., 34 optimal responses out of 68). All tests were significant at a significance level well under 0.001. Since computer performance is optimal with respect to the four cues available to the computer, the explanation for superior human performance (leaving aside the very remote possibility of chance) must involve use by humans of the extra cue. Participants were successful in learning the value of the extra cue and in employing that knowledge to improve overall human-computer performance.

The next step was to compare the six experimental conditions among themselves. A two-factors repeated-measures analysis of variance was performed on the data, the results of which are summarized Table 2-5.

Table 2-5

Overall Analysis of Variance

Source	df	Sum of Squares	Mean Square	F-ratio	p
Allocation	2	187.0625	93.53125	1.310146	>.25
Workload	1	880.125	880.125	15.60896	<.001
Allocation × Workload	2	392.0313	196.0156	5.294706	<.006
Allocation × Subjects	46	3283.938	71.38995		
Workload × Subjects	23	1296.875	56.38587		
Allocation × Workload × Subjects	46	1702.969	37.02106		

The effect of task allocation was not significant. However, the impact of workload was highly significant ($\underline{F}(1,23) = 15.61$, $p<.001$), and the task

allocation by workload interaction was also highly significant ($F(2,46)$ = 5.29, p=.006).

To explore further the interaction between task allocation and workload, a subsidiary analysis was performed. Separate ANOVAs were used to test the impact of task allocation at each of the two levels of workload. The results are shown in Table 2-6. Task allocation had no effect at low workload, but had a highly significant effect ($F(2,46)$ = 5.23, p=.006) under high workload.

Table 2-6

Subsidiary Analysis of Variance

Workload Condition	Source	df	Sum of Squares	Mean Square	F-ratio	p
Low	Allocation	2	51.85938	25.92969	.44704	>.50
Low	Allocation × Subjects	46	2668.141	58.00306		
High	Allocation	2	527.25	263.625	5.229865	.006
High	Allocation × Subjects	46	2318.75	50.40761		

In order to obtain a more detailed understanding of these data, a series of contrasts involving paired comparison t-tests was carried out. These comparisons resulted in findings that fit a definite pattern: (1) at low levels of workload, the three allocation conditions were equivalent; (2) high workload caused decrements in the manual and override conditions, but had no effect on performance in the screening condition. Thus, performance under the conditions (manual/low-load), (screening/low-load), (override/low-load), and (screening/high-load) were all the same; but marked worsening in performance occurred under the two treatment combinations (manual/high-load) and (override/high-load). Table 2-7 graphically illustrates this pattern. Shaded calls were found to be essentially equivalent in performance.

Table 2-7

Equivalent Conditions (Shaded)

	MANUAL	SCREENING	OVERRIDE
LOW LOAD	(shaded)	(shaded)	(shaded)
HIGH LOAD		(shaded)	

The presence of this pattern was, finally, tested directly. We computed the average performance under the four shaded conditions for each individual, and subtracted the average performance under the other two conditions. For these data, the t-statistic was 5.13--more significant than any

other paired comparison we have considered above, at far below the $p<.001$ level. We concluded that, under manual and override conditions, a real worsening of performance occurred under high load, but that for the screening condition, load did not affect performance within the range tested.

In summary, within the limitations of the air defense simulation employed, it was clear that major human contributions center around human abilities to perceive patterns, to learn, and to adapt over relatively short periods of time. Although the prototype representation of the air defense system employed here was much simpler than any real system, the results suggest that some strategies for task allocation will work better than others and that in some settings under high processing loads, neither fully manual nor fully automated systems work as well as a scheme permitting some collaboration between the system's human and computer elements.

Under low-workload conditions, and where the computer model in the ID task was significantly incomplete, any of the three conditions involving human participation was superior to computer-only performance. However, at high workload this advantage could only be maintained if the computer assisted the human by directing the user's attention to subproblems where his/her contribution was most needed. In short, a complementarity has been observed in which the human helps the computer by learning and adapting to novel situations, and in which the computer helps the human by reducing the size of the problem.

2.2 Additional Analyses and Cognitive Modeling

The analyses performed in Phase I (and summarized within Section 2.1.2 of this report) focused only on the primary conflict condition, i.e., the condition where two cues pointed toward hostile and two cues toward friend. This condition was of primary concern because it was clearly the condition where Ps should use the "extra cue," which was not available to the computer. In short, it was the case of maximum ID conflict for the automated system and, therefore, where the human's collaboration could have the biggest impact. However, the extra cue also could help in cases where one cue pointed one way and three cues the other. We were clearly interested in how Ps responded in the lesser conflict case and more generally for all cue combinations, for the different allocation and workload conditions. The purpose of this broadened analysis was to shed some light on what information-processing strategies Ps were, in fact, using in making their judgments.

2.2.1 Information-processing strategies. To focus our discussion of the data, it will help to consider two alternative information-processing strategies which participants might have employed in the manual condition. Figure 2-2a is an English description of an "optimal" approach to the task. Recall that the participant received information on two computer screens: a graphics screen containing the "extra" cue and a single additional graphic cue, and an alphanumeric screen containing three alphanumeric cues. Since the extra cue was more diagnostic than any three of the other cues combined, the optimal strategy involved looking at the graphics screen first. If both or neither of the cues on that screen were hostile, the

participant could make a decision on that target (to shoot or not shoot, respectively) without consulting the alphanumeric screen at all. If the participant does have to switch his attention to the other screen, the optimal strategy is to scan the alphanumeric cues searching for one that confirms the extra cue (whether hostile or friendly); if one is found, any remaining cues need not be examined.

Look at Extra Cue and Graphic Cue on Graphics Screen.

If both cues are hostile, shoot then go to next target.

If neither cue is hostile, go to next target.

If extra cue only is hostile,
switch to alphanumeric screen.
Look at cues until one additional hostile cue is found, then shoot, switch screens, and go to next target.
If no additional hostile cue is found, switch screens and go to next target.

If graphic cue only is hostile,
switch to alphanumeric screen.
Look at cues until one additional friendly cue is found, then switch to graphics screen and go to next target.
If no additional friendly cue is found, shoot, then switch screens and go to next target.

Figure 2-2a. Optimal processing strategy for manual condition.

Figure 2-2b presents an alternative, simpler strategy which does not take into account the differential diagnosticity of the cues. On the "Majority of Confirming Dimensions" (MCD) strategy, a majority vote of the five cues is sufficient to determine target ID. In this case, participants would be expected to start with the alphanumeric rather than the graphics screen, since it has more cues. If all 3 alphanumeric cues are hostile or friendly, participants need not switch attention to the graphics screen at all, since 3 is already a majority. Otherwise, if 2 alphanumeric cues are hostile or 2 cues are friendly, the best approach is to scan the graphics cues searching for one that confirms the predominant vote of the 3 alphanumeric cues; if the first graphic cue looked at confirms this vote, the remaining cue need not be examined.

Look at All Cues on Alphanumeric screen.

If 3 cues are hostile, shoot and go to next target.

If 2 cues are hostile,
switch to graphics screen.
Look at cues until one additional hostile cue is found, then shoot, switch back to alphanumeric screen, and go to next target.
If no additional hostile cues are found, switch screens and go to next target.

If 1 cue is hostile,
switch to graphics screen.
Look at cues until one friendly cue is found, then switch screens and go to next target.
If both cues are hostile, shoot, then switch screens and go to next target.

If no cues are hostile, go to next target.

Figure 2-2b. Majority of confirming dimensions strategy for manual condition.

We do not believe that any participant followed either of these strategies in its purest form. However, they provide an extremely useful framework for making sense of the data to be described below. More generally, we hope to illustrate how different human-computer task allocation schemes can be understood in terms of their impact on user cognitive strategies.

Figures 2-3a and b give a more precise specification of each strategy in terms of elementary information processes. (Note: Screen 1 is the alphanumeric screen; screen 2 is the graphics screen.) It is impressive that these strategies can be analyzed into rules which are entirely composed of *seven* elementary operations (though we should stress that each of these operations could be decomposed further into still more "elementary" constituents):

- Go_to (a screen)
- Locate (a target)
- Read (a cue value)
- Categorize (a cue value as friendly or hostile)
- Add (1 to a running count)
- "Shoot" - type (target number)
- Set (a goal, i.e., all targets or all cues).

```
GO_TO (SCREEN-2)
SET (TARGET)
LOCATE (TARGET)
READ (VALUE OF EXTRA CUE)
CATEGORIZE (VALUE OF EXTRA CUE)
READ (VALUE OF GRAPHIC CUE)
CATEGORIZE (VALUE OF GRAPHIC CUE)
```

IF CATEGORY (EXTRA CUE) =

IF CATEGORY (GRAPHIC CUE) =	HOSTILE	FRIENDLY
HOSTILE	SHOOT (TARGET) NEXT (TARGET)	GO_TO (SCREEN-1) [A]
FRIENDLY	GO_TO (SCREEN-2) [B]	NEXT (TARGET)

```
[A] LOCATE (TARGET)
    SET (ALPHANUMERIC CUE)
    READ (CUE VALUE)
    CATEGORIZE (CUE VALUE)
    IF CATEGORY (CUE) = FRIENDLY THEN
        GO_TO (SCREEN-2)
        NEXT (TARGET)
    NEXT (ALPHANUMERIC CUE)
    SHOOT (TARGET)
    GO_TO (SCREEN-2)
    NEXT (TARGET)
```

```
[B] LOCATE (TARGET)
    SET (ALPHANUMERIC CUE)
    READ (CUE VALUE)
    CATEGORIZE (CUE VALUE)
    IF CATEGORY (CUE) = HOSTILE THEN
        SHOOT (TARGET)
        GO_TO (SCREEN-2)
        NEXT (TARGET)
    NEXT (ALPHANUMERIC CUE)
    GO_TO (SCREEN-2)
    NEXT (TARGET)
```

Figure 2-3a. "Optimal" processing strategy for manual condition.

```
GO_TO (SCREEN-1)

SET (TARGET)

LOCATE (TARGET)

SET (ALPHANUMERIC CUE)

READ (CUE VALUE)

CATEGORIZE (CUE VALUE)

IF CATEGORY (CUE) = HOSTILE THEN ADD (COUNT, 1, COUNT)

NEXT (ALPHANUMERIC CUE)

IF COUNT = 3 THEN
     SHOOT (TARGET)
     NEXT (TARGET)

IF COUNT = 2 THEN
     GO_TO (SCREEN-2)
     LOCATE (TARGET)
     READ (VALUE OF EXTRA CUE)
     CATEGORIZE (VALUE OF EXTRA CUE)
     IF CATEGORY (EXTRA CUE) = HOSTILE THEN
          SHOOT (TARGET)
          GO_TO (SCREEN-1)
          NEXT (TARGET)
     READ (VALUE OF GRAPHIC CUE)
     CATEGORIZE (VALUE OF GRAPHIC CUE)
     IF CATEGORY (GRAPHIC CUE) = HOSTILE THEN SHOOT (TARGET)
     GO_TO (SCREEN-1)
     NEXT (TARGET)

IF COUNT = 1 THEN
     GO_TO (SCREEN-2)
     LOCATE (TARGET)
     READ (VALUE OF EXTRA CUE)
     CATEGORIZE (VALUE OF EXTRA CUE)
     IF CATEGORY (EXTRA CUE) = HOSTILE THEN
          READ (VALUE OF GRAPHIC CUE)
          CATEGORIZE (VALUE OF GRAPHIC CUE)
          IF CATEGORY (GRAPHIC CUE) = HOSTILE THEN SHOOT
               (TARGET)
     GO_TO (SCREEN-1)
     NEXT (TARGET)

IF COUNT = 0 THEN NEXT (TARGET)
```

Figure 2-3b. "Majority of confirming dimensions" strategy for manual condition.

Specifications of this kind can provide the basis for simulations of human-computer performance under various task-allocation conditions, and prediction of human workload in terms of the number of required operations. The data analysis to be described here, however, does not rely on a level of detail beyond the English descriptions in Figures 2-2a and b.

Let us briefly consider where these two strategies might diverge in their implications for the manual condition data.

- *ID Performance*. MCD strategy predicts 3 hostile cues of any kind are sufficient for hostile ID; optimal strategy predicts extra cue plus at least one other cue, or all four of the other cues, would be required.

- *Cue Reliance*. Optimal strategy implies greater reliance on extra and graphic cues; MCD strategy implies greater reliance on alphanumeric cues.

- *Latencies*. No divergence (for analyses performed). Both strategies predict decreasing latency with increasing number of hostile cues.

A principal goal of our analysis is to explore the impact of the two independent variables--workload and allocation scheme--on the selection and execution of information-processing strategies. For example, increasing workload might have any of the following effects:

- use of fewer cues per target;

- increase in performance errors (e.g., misreading cue values);

- qualitative change in processing approach (for each target);

- increase in the response criterion for a hostile ID; or

- analysis of fewer targets.

By the same token, providing a facility for screening might protect performance from decrements due to high workload by:

- permitting use of more or better cues per target;

- reducing performance errors;

- facilitating more optimal adjustments in strategy;

- permitting more optimal response criteria; or

- focusing attention on targets most in need of human analysis.

Thus, the cognitive modeling methodology to be explored here holds the promise of clarifying and explaining the results obtained in the

earlier analysis, and pointing the way to improved design of human-computer allocation schemes.

2.2.2 Manual condition.

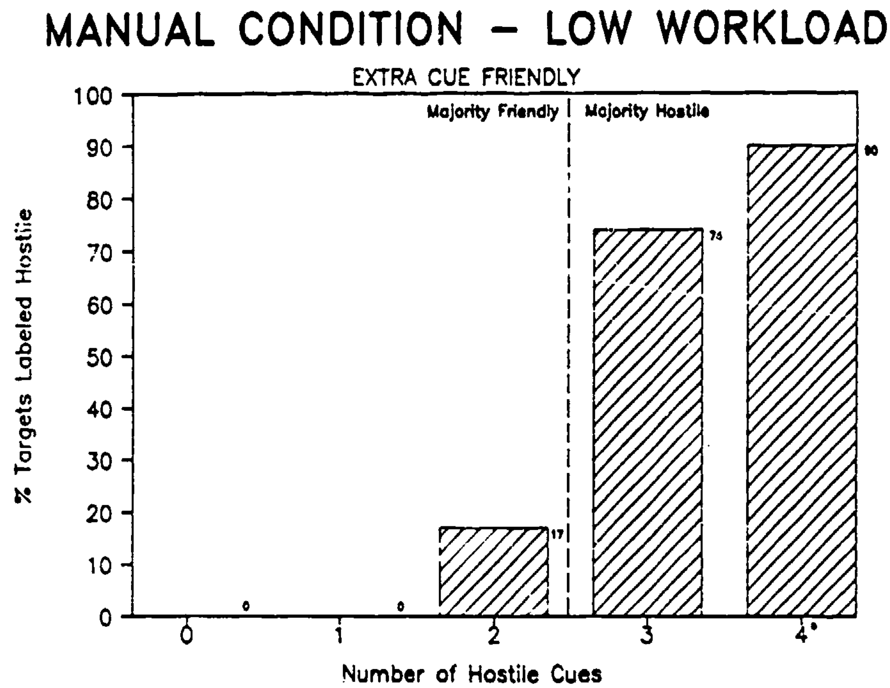
ID Performance Data. Figure 2-4 shows the percentage of targets labeled hostile in the manual condition under low workload, as a function of the total number of hostile cues, excluding the extra cue. In the top graph, the extra cue was friendly; in the lower graph, the extra cue was hostile. (Recall that in the manual condition, the target was always represented by the symbol "U" for unknown. The participant typed the target's number to identify it as "hostile;" targets which remained unknown were, in effect, treated as "friends.")

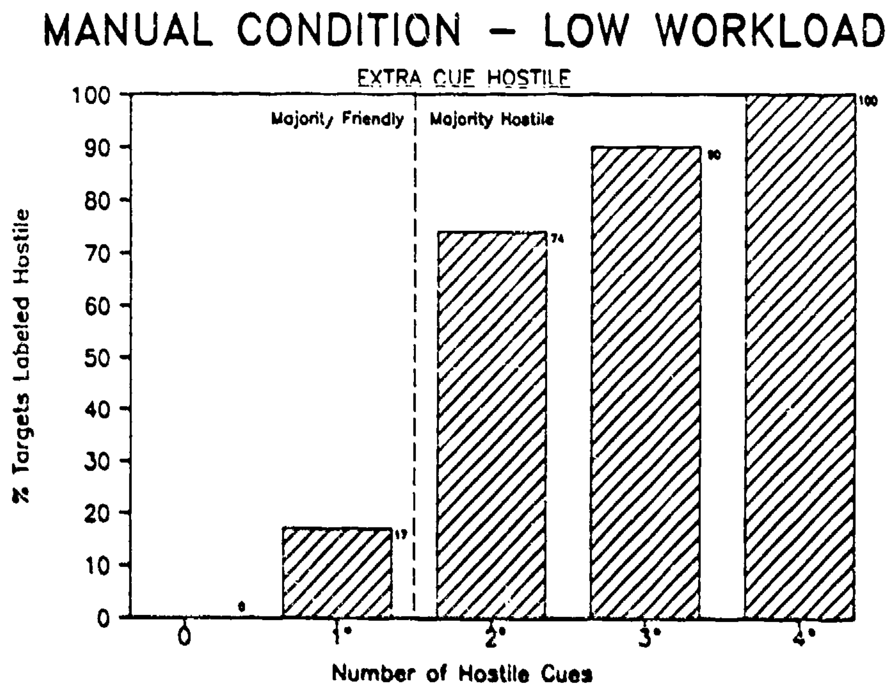
What is most noteworthy here is (a) in the top figure, the large difference between responding when only 2 cues were hostile and responding when 3 cues were hostile; and (b) in the bottom figure, the similar large difference between responding when 1 cue was hostile (in addition to the extra cue) and responding when 2 cues were hostile (in addition to the extra cue). The two charts support the MCD strategy: they suggest that the extra cue was like any other in perceived diagnosticity, and that participants (on average at least) responded with a hostile ID when 3 out of the 5 cues were hostile. By contrast, the optimal strategy implies 100% responding in the upper chart only when 4 cues are hostile, and 0% elsewhere; and 100% responding in the lower chart for 1, 2, 3, and 4 cues hostile in addition to the extra cue.

A second observation qualifies the above conclusion: despite the large jumps just noted, in both charts the chance of a target's being labeled hostile increases to some extent across the board with the number of hostile cues. It does not go from 0% to 100% when the number of hostile cues reaches a majority. A variety of explanations are consistent with this finding: e.g., (a) a probabilistic response rule according to which, for example, when cues suggest the probability of hostile is X%, participants respond "hostile" X% of the time (a "probability matching" strategy); (b) errors in evaluating cues, i.e., occasionally misreading hostile cues as friendly and friendly cues as hostile. Probability matching does not easily explain the large jumps in responding when the number of hostile cues reaches a majority. Occasional misreading of cues is certainly a possibility; however, the error frequency that would be required to account for these data seems implausibly high, especially in the low-workload condition.

For example, to explain the 21% response rate when only 2 cues are hostile (averaging over extra cue status), we must suppose that one or more of the 3 friendly cues are misread as hostile for 21% of these targets. To explain a 94% response rate when 4 cues are hostile, it must be assumed that at least *two* of the four hostile cues are simultaneously misread as friendly for 6% of these targets.

It is worth considering an explanation which flows more naturally from the MCD strategy, and which also accounts easily for the probabilistic effects. A given participant may on occasion utilize only a subset of the





- = Predicted response if 43% of the time on' 3 cues are sampled

* = Responding should be 100% on optimal strategy, 0% elsewhere

Figure 2-4. Percentage of targets labeled hostile as function of number of hostile cues (manual, low workload)

available cues; for example, one may look at 3 of the 5 available cues. On those occasions, a hostile response will occur when a majority of the *utilized* cues is hostile. Table 2-8 gives the probability that a majority of the sampled subset will be hostile for varying numbers of hostile cues in the total set, and for varying sizes of the sample. (Predicted values are based on the hypergeometric distribution; Feller (1950).

Table 2-8

Probability a Majority of Sampled Cues are Hostile

Number of Hostile Cues in Total Set of 5		Size of Sampled Set 1	2*	3	4*	5
Majority Hostile	5	1.0	1.0	1.0	1.0	1.0
	4	.8	.6	1.0	1.0	1.0
	3	.6	.3	.7	.4	1.0
Majority Friendly	2	.4	.1	.3	0	0
	1	.2	0	0	0	0
	0	0	0	0	0	0

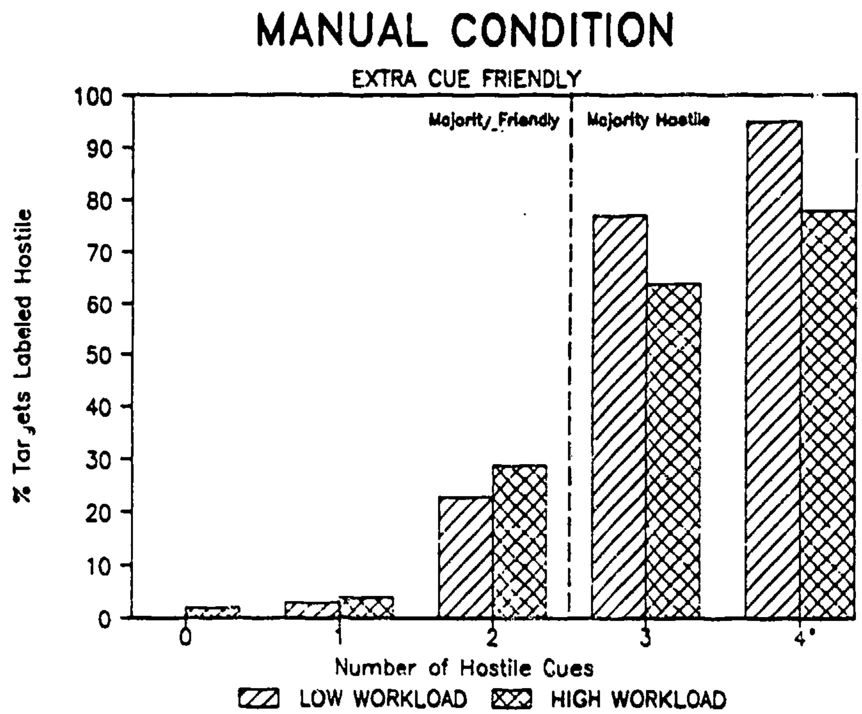
* A majority is defined for even sample sizes as *more* than half.

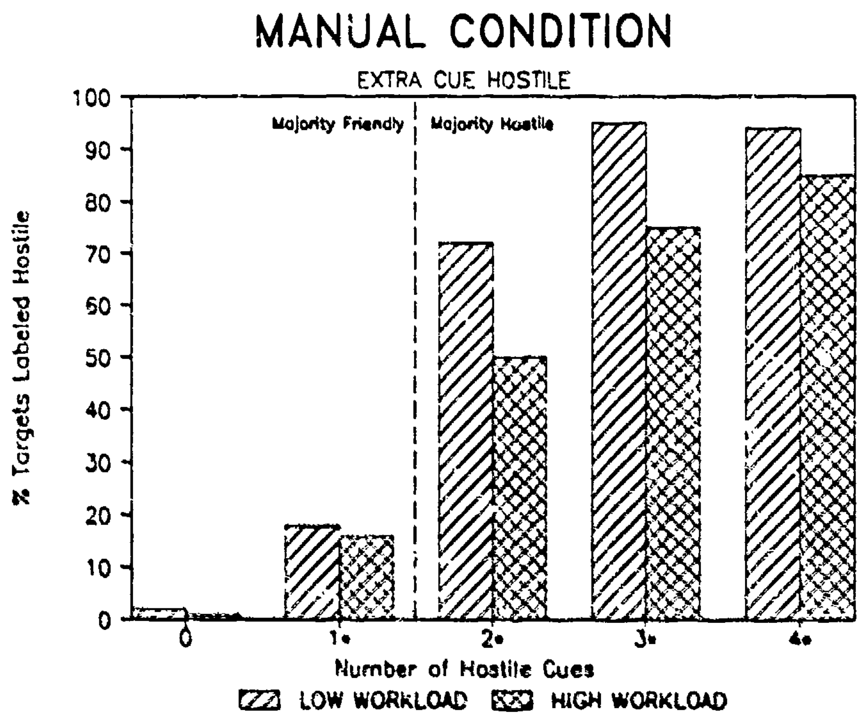
Under this strategy, then, even when the total set of cues does *not* have a hostile majority, a sampled subset of cues may have a hostile majority; and even when the total set of cues *does* have a hostile majority, the sampled subset may not. The larger the proportion of the total cues which are hostile, the more likely it is that a majority of the selected subset will be hostile. Therefore, the probability of a hostile response increases with the total number of hostile cues, but it is not as high as if all cues were sampled.

Note that *all* sampling strategies imply a probability of responding of 1.0 when all 5 cues are hostile. Yet in that condition, participants responded only 93.3% of the time. It is therefore plausible to assume that about 6% of the targets were not attended, even under conditions of low workload.

Figure 2-5 presents the ID performance data for both low and high workload in the manual condition. High workload shows a similar pattern to low workload, with the largest jumps in responding occurring where the number of hostile cues reaches a majority (whether the extra cue was hostile or friendly). In high workload, as expected, fewer targets were classified hostile. The most striking observation, however, is that this discrepancy between high- and low-workload performance only occurs when the number of hostile cues is a majority, i.e., at 3 and 4 on the upper chart, and at 2, 3, and 4 on the lower chart.

Several possible explanations of the workload effect can be dismissed. First, these findings cannot be accounted for merely by assuming that participants examine fewer targets under high workload, but use essentially the same strategy for the targets that are examined. In that case,





* = Responding should be 100% on optimal strategy, 0% elsewhere

Figure 2-5. Performance in manual condition for low and high workload

the reduction in responding would be a *consistent* proportion across all the different numbers of hostile cues, rather than concentrated at the conditions where hostile cues are in the majority. Second, the findings cannot be accounted for by supposing that participants adopted a different response criterion, since the high workload data continue to support the predominant use of 3 hostile cues as a criterion for a hostile ID.

Finally, an account in terms of increased errors in cue evaluation seems implausible. To account for the 57% average response rate when 3 cues were hostile, we must imagine that one or more hostile cues was misread as friendly for 43% of the targets. To account for the 77% response rate when 4 cues were hostile, we must suppose that for 23% of the targets at least *two* hostile cues were misread simultaneously. In addition, an account in terms of cue-evaluation errors does not easily account for the asymmetric effect of workload on targets which have a majority of hostile cues and targets which do not. It must be assumed, rather arbitrarily, that workload affects the tendency to regard hostile cues as friendly, but not vice versa.

The workload effect can be accounted for more naturally in terms of a change in participants' cue-sampling strategy: an increased tendency under high workload to utilize only a subset of the cues for each target. This hypothesis nicely captures an important qualitative feature of the data: hostile ID responses *decreased* under high workload when there was, in fact, a majority of hostile cues, but was relatively unaffected when there was not a true majority for hostile.

Cue Dependence Data. Some confirmation for these hypotheses, and a more detailed picture of the way participants utilized specific cues to arrive at ID decisions, is provided by regressing the ID decisions against cue values. Figure 2-6 shows the resulting regression coefficients, for both low and high workload, in the manual condition. The MCD strategy predicts that participants will use the graphics screen only when the alphanumeric screen is inconclusive; hence, there will be greater reliance on the alphanumeric cues (an1, an2, an3) in comparison to the extra cue and the graphic cue (g1); the optimal strategy predicts the converse. As shown in Figure 2-6, the coefficients for the alphanumeric cues are significantly higher than for the extra and graphic cues, supporting the MCD model.

Dependence on cues of all types was less under high workload than under low workload. Nevertheless, use of the extra and graphic cues was significantly more reduced than the use of alphanumeric cues. This is consistent with our hypothesis that under high workload, participants utilize fewer cues for each target; specifically, it suggests that participants are less likely under high workload to switch their attention to the second (graphics) screen even when the alphanumeric screen is inconclusive.

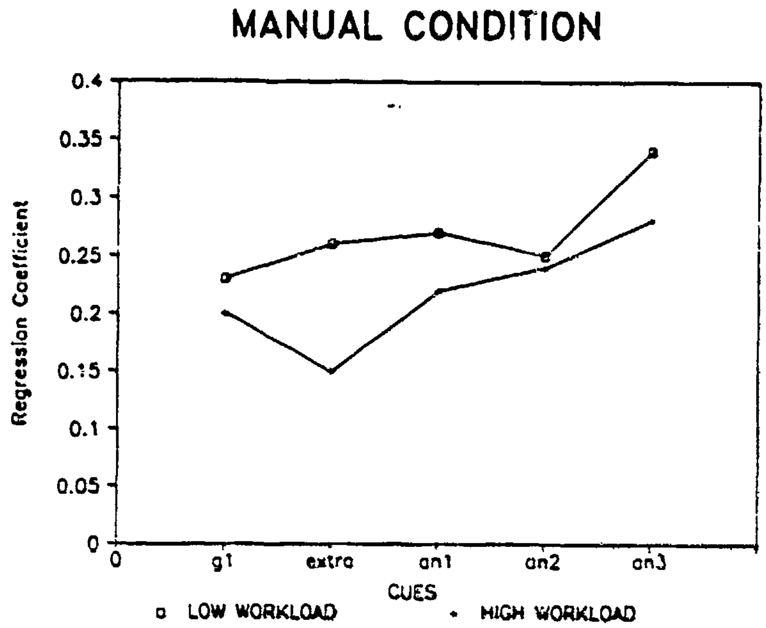


Figure 2-6. Regression coefficients against cue values.

These conclusions are further confirmed by Table 2-9, which shows the percentage of targets labeled hostile in the manual condition as a function of how many alphanumeric and graphics-screen cues were hostile. There are linear trends (significant) for both alphanumeric and graphics-screen cues in high- and low-workload conditions (although the effect is weaker in high workload). Two findings are of interest. First, when three cues are hostile instead of two, regardless of combination, there is a big jump (roughly 50%) in the frequency of responding. This suggests a majority principle is operating. Secondly, as noted in the regression, the alphanumeric cues carry a bigger weight than the graphics-screen cues. For example, if 3 alphanumeric and no graphics-screen cues are hostile, the percentage of firing is 82.30% (in the low-load condition) compared to 75.69% when 2 AN cues and 1 graphics-screen cue are hostile compared to 70.33% when 1 AN cue and 2 graphics-screen cues are hostile. This pattern holds throughout the matrix.

2.2.3 Screening condition. In the screening condition, again three hostile cues seemed sufficient for classification as hostile. Here the data are more difficult to analyze, however, because the computer is making classifications in all but the 2 vs. 2 condition. To facilitate this analysis, cell-by-cell comparisons are made between the screening conditions and the manual conditions, which can be viewed as baseline, unaided conditions. Table 2-10 indicates the computer symbols, and required operator responses to change the symbols, in the different conditions.

When the 5th cue is on the friendly side and there are 0 and 1 hostile cues, participants behaved the same in the screening conditions as in the manual conditions. This could be attributable either to participants not feeling the number of cues was sufficient to change the computer's ID

Table 2-9

Percent Engaging of Target by Cue Type

Manual Condition - Low Workload

# of Graphics Cues Hostile	# of Alphanumeric Cues Hostile 0	1	2	3
0	.42%	3.36%	32.49%	82.30%
1	2.74%	16.03%	75.69%	94.89%
2	23.28%	70.33%	93.97%	93.33%

Manual Condition - High Workload

# of Graphics Cues Hostile	# of Alphanumeric Cues Hostile 0	1	2	3
0	1.67%	3.87%	33.97%	60.19%
1	3.74%	18.46%	60.85%	77.13%
2	16.18%	43.52%	75.75%	85.83%

Above dotted line: minority of cues are hostile.
Below dotted line: majority of cues are hostile.

Table 2-10

Computer Symbols and Required Operator Responses to Change Symbols in the Different Interface Conditions

Interface Condition	Computer Symbol	Response Requirement
Manual	?	Change only to h
Screening	f	Change to h (or leave as f)
	?	Assign
	h	Change to f (or leave as h)
Override	f	Change to h (or leave as f)
	h	Change to f (or leave as h)

or to participants ignoring those cases (since they are only prompted when the condition is 2 vs. 2). One can try to distinguish between these explanations by examining the 3 hostile cue situation when the extra cue is friendly. The optimal response is "friend." Keep in mind the computer has classified this target as hostile, unlike the manual condition where the default classification is friendly.

In the screening, low-load condition with 3 hostile cues (5th cue friendly), participants changed the ID 21.46% of the time, which means they considered 21.46% of these targets friendly and 78.54% of them hostile. In the analogous manual condition, they considered 23.54% of the targets to be friendly and 76.54% of them to be hostile. The numbers for these two conditions are virtually identical, suggesting that participants are looking at cues rather than just ignoring those targets. The data for the 4 hostile cues are also comparable, 13.89% judged friendly in the screening condition and 6.25% judged friendly in the manual condition.

In the screening, high-load condition with 3 hostile cues (5th cue friendly), participants changed the ID 13.13% of the time (thus judging 13.13% of the targets friendly), while in the manual, high-load condition, they called 63.75% of the targets hostile, thus judging 36.25% of the targets to be friendly. The finding that more targets are judged friendly using the same cues in the high manual condition compared to the high screening condition suggests that in the high load, screening condition participants are paying less attention to the non-prompted targets. The argument is that if participants were paying as much attention in the screening condition as they were in the manual condition, they would have changed the computer's ID 36% of the time rather than 13% of the time.

We can gain further support for this "attention" hypothesis by looking at the 4 hostile cue conditions. When the extra cue points toward friend, participants judge 11.81% percent of the targets friendly (based on "change" scores) in the screening, high-load condition; whereas in the manual, high-load condition, they judge 21.52% friendly. Again, the finding [illegible] participants are overriding less in the screening condition than

would be expected based on the manual condition's data (the number might have been still lower if not for floor effects) suggests that they are paying less attention to these cases.

When the 5th cue is on the hostile side, the pattern of results confirmed our previous assertions. Again, in the 0 and 1 hostile cue conditions for both high and low workload, the results seem the same for screening and manual conditions: participants judged targets as predominantly friendly.

When the cues are 2 vs. 2, an interesting pattern emerges. First, for the low-load conditions, when the extra cue is hostile, the number of targets judged hostile is the same across the manual and screening conditions (73.41 versus 77.45, respectively). However, when we look at the high-load conditions, we find that the participants judged more targets as hostile in the screening condition (69.61) than they did in the manual condition (50.38). Since the participants were more accurate in the screening condition than they were in the manual condition, this suggests that they were paying more attention to the 2 vs. 2 condition in that condition. This suggests an attentional-focusing heuristic. This pattern of results did not, however, appear when the extra cue was friendly.

To further validate the "attention" hypothesis, we compare the 3 and 4 hostile cue conditions when the extra cue is hostile. In the low-load, screening condition, the number of targets judged friendly were 6.81% and 8.75% for the 3 and 4 hostile cue conditions, respectively. This was comparable to the analogous manual conditions, where the number of targets judged friendly were 5.55% and 6.67%, respectively. In the high-load screening conditions, the number of targets judged friendly in the 3 and 4 hostile cue conditions were 9.44% and 8.33%, respectively. The analogous manual conditions were 24.17% and 14.17%. This discrepancy suggests that participants in the screening, high-load condition were paying less attention to these cue combinations than in the manual, high-load condition, since they overrode less frequently than would be expected if they had paid an equal amount of attention.

To summarize the screening conditions, participants in the low load conditions responded similarly to those in the manual, low-load conditions. This suggests that they were not just focusing on the the 2 vs. 2 cue conditions but rather were focusing on other conditions as well. In the high-load conditions, participants compensated for the high load by focusing mostly on the 2 vs. 2 condition where they were prompted.

2.2.4 Override condition. Finally, we look at the override conditions. The first thing to note is that the percentages are low, suggesting that little overriding occurred. In general, the numbers were lower for the high-load conditions than they were for the low-load conditions, suggesting again that participants compensated for increased load by ignoring some of the targets. This is especially evident in the 2 vs. 2 cue conditions. In this condition, the computer randomly assigned targets as friend or foe. Hence, we expect 50% accuracy and that 50% of the time participants should change the ID if they are paying attention to the cues. We see that in the low-load conditions, participants changed the ID roughly

42% of the time (across both 5th cue conditions) and in the high-load conditions about 30% of the time (across both 5th cue conditions). What is interesting here is that the difference between high- and low-load conditions is greatest in the 2 vs. 2 conditions, suggesting that high load degrades people's ability to make judgments in the most ambiguous conditions.

We now compare the override conditions to the screening and manual conditions. What seems to characterize the override conditions (both high and low load) is that participants did virtually all their changes in 2 vs. 2 and 3 cue conditions, i.e., in the cases of (a) 3 hostile cues plus "extra cue" friendly, and (b) one cue hostile plus the "extra cue" hostile (3 friendly cues). While this was to some extent true in the screening conditions, there the emphasis was much more on the 2 vs. 2 conditions.

These results suggest a general continuum of attentional focus. In the manual conditions, participants tried to focus on all cue conditions and hence there is a greater difference in high vs. low load conditions. The override conditions focused much more on 2 vs. 3 cue conditions (including the extra cue), hence a smaller difference in overriding across load conditions. Finally, in the screening conditions, the focus was primarily on the 2 vs. 2 condition (excluding the extra cue) and, hence, the amount of ID changes is most similar across high and low load conditions.

To give an overall summary, in low-load conditions, the amount of changes were about the same across manual, screening and override conditions, except the latter had a slight tendency to override only in 2 vs. 3 cue conditions. Also, differential changes seem to occur across cue conditions suggesting that participants were looking at all the cues.

In high-load conditions, overriding drops off, suggesting that people look at fewer than all the targets. The results of the screening and override conditions suggest that people look for heuristics to do this. In the screening case, they focused largely on the 2 vs. 2 cue cases prompted by the computer. In the override case, they focused on cases of maximum ambiguity 2 vs. 3 cue cases. In fact, Ps were actually somewhat more accurate for the minor "conflict" targets, i.e., for those targets where 3 cues pointed one way but the extra cue pointed another, in the override than screening condition under both levels of workload. The "attention focusing" hypothesis provides a reasonable explanation for this finding.

In the screening condition, as expected, participants rely more on the extra cue, especially in the 2 vs. 2 cue condition where the computer prompts the user with a "?". In fact, in the 2 vs. 2 condition, the extra cue is the only one with a significant coefficient.

Comparing the screening condition to the manual condition, we see that the coefficients are far less stable and lower (except for the extra cue in the 2 vs. 2 condition), suggesting that when participants are expecting the computer to make a decision and prompt them when it is uncertain, they are highly unsystematic in the strategies they use to decide

whether or not to update the computer's decision. Moreover, except in the 2 vs. 2 condition, the R^2s are extremely low.

2.2.5 Response latencies. The analyses below focus on participants' response latencies (i.e., how quickly they responded), and how they were influenced by workload, the computer's identification of the target in the screening and override conditions, and also which half of the session the participant was in. The results of these analyses are presented in Tables 2-11, 2-12, and 2-13 for the manual, screening, and override conditions, respectively. The one observation that was overwhelming for these analyses was that participants responded faster under conditions of low workload compared to high workload. The effect size was on the order of 10-15 seconds. This finding is counter-intuitive and at odds with the findings of other researchers (e.g., Payne, Bettman, and Johnson, 1986). One possible explanation for it that is consistent with the results presented earlier in this report is that, in an effort to cope with greater workload, participants spend more time examining those targets they choose to examine and, thereby, attempt to minimize the likelihood of mistakes caused by high workload. This explanation is, however, post-hoc and in need of further investigation in Phase II.

As for target ID, targets identified as hostile by the computer in the override and screening conditions were responded to faster than targets identified as friendly. Targets labeled as unknown or "?" in the screening condition fell somewhere in the middle. An interesting parallel finding relates the size of the regression coefficients for the cues to whether or not participants predicted the target was hostile and friendly. In both override and screening conditions, there is a general tendency for more cues to have significant regression coefficients and for the coefficients to be larger in general when targets are labeled friendly by the computer than when they are labeled hostile. When taken with the response latencies analysis, this suggests that when a target is labeled friendly, participants take longer to respond and seem to rely more heavily on more cues than when a target is labeled hostile. This may be interpreted as a difference in perceived utility for own and enemy craft, i.e., own craft is worth a lot so participants take more time and look at more evidence before shooting at a target labeled friendly, but are more willing to let potentially hostile craft through based on a quicker decision using the cues less. It may be interesting if this effect holds up with real air defense participants; if so, it may be analogous to previous research (e.g., Cohen, Brown, and Chinnis, 1986) where target importance enters into the decision.

Table 2-11

Response Latency, Manual Condition

	Mean Response Latency in Seconds	
	1st Half	2nd Half
Low Load	16.229	15.839
High Load	31.816	29.103

For Load	F(1,4230) = 870.307, p = .000
Half	F(1,4230) = 10.576, p = .001
Load & Half	F(1,4230) = 5.330, p = .020

Table 2-12

Response Latency, Screening Condition

	Mean Reponse Latency in Seconds			
	High Workload		Low Workload	
	1st Half	2nd Half	1st Half	2nd Half
ID-friendly	34.275	30.872	21.486	23.432
ID-"?" (2 vs. 2)	32.617	30.904	18.887	19.708
ID-hostile	29.696	25.596	15.346	17.708

There are three general trends:

1) Low workload is faster than high workload (F1,2289) = 156.898, p = .000.

2) ID = hostile is faster than ID = "?", F(1,2289) = 14.800, p = .000 and ID = "?" is faster than ID = friendly, F(1,2289) = 3.440, p = .060.

3) Under high workload, subjects get faster going from 1st half to 2nd half; under low workload, subjects get slower going from 1st half to 2nd half, F(1,2289) = 7.109, p = .000.

Table 2-13

Response Latency, Override Condition

	Mean Response Latency in Seconds			
	High Load		Low Load	
	1st Half	2nd Half	1st Half	2nd Half
ID=friendly	30.564	28.425	19.802	20.592
ID=hostile	28.753	27.163	18.264	18.743

There are three general trends:

1) Low workload is faster than high workload, $F(1,1782) = 185.534$, $p = .000$.

2) ID = hostile is faster than ID = friendly, $F(1,1782) = 5.535$, $p = .018$.

3) Under high load, subjects get faster going from 1st half to 2nd half; under low load, subjects get slower going from 1st half to 2nd half, $F(1,1782) = 3.313$, $p = .065$.

Another finding of interest with regard to the latency data has to do with the effect of session half. In the manual condition, participants' speed increases in the second half of the session. However, this effect is largely obtained in the high-load conditions. In the screening and override conditions, an interesting effect occurs. While there is no main effect for session half, there is an interaction effect with workload, i.e., participants get faster going from 1st half to 2nd half, under high workload, but actually get *slower* going from 1st half to 2nd half under lower workload.

One possible hypothesis for explaining this finding is that, in low workload, participants tend to abandon strategies going from the first half to second half of the session. This suggests that as the session progresses, participants start looking at more cues which could result in a longer response latency. If we had latency measures of non-override responses, which we do not, we could see whether this effect held up there as well. As such, this hypothesis is purely speculative.

The final finding of interest focuses on the response latencies for participants in the manual condition as a function of the number of cues pointing to the target being hostile. Figure 2-7 shows that the response latency decreased with the number of hostile cues up until three, and then basically leveled off. This supports an "MCD" strategy; however, given conflict among the cues, an "optimal strategy" can not be ruled out because the latter strategy suggests a sequential processing of cues (3, 4, or 5) until a threshold decision point is reached. In contrast, the response latency is roughly constant with the number of cues in the high-workload condition. One possible explanation that is consistent with the "attention

focusing" hypothesis is that, under high workload, participants are more likely to emphasize a target focus and, therefore, look at all cues for each target they select, even if they are using an "MCD" strategy. Although the explanations are speculative, they are consistent with the finding of significant regression coefficients on all five cues in both workload conditions.

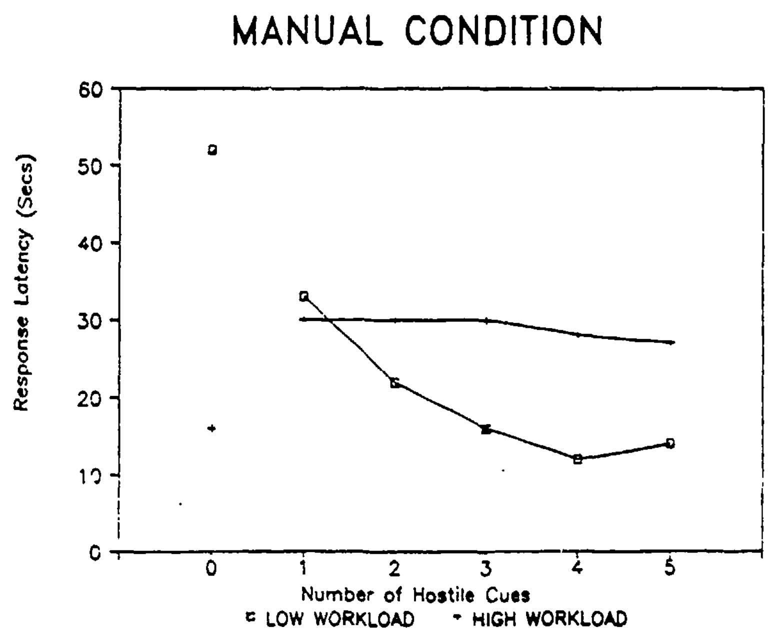


Figure 2-7. Mean response latencies for participants in the manual condition as a function of the number of hostile cues.

3.0 PHASE II EXPERIMENTS WITH ACTIVE-DUTY AIR DEFENSE OPERATORS

This section of the report describes controlled experiments testing the relative effectiveness of alternative human-machine allocation schemes with active-duty, U.S. Army air defense operators. Section 3.1 presents an overview of the basic information-processing principles guiding our research, and the experimental designs for the two experiments testing these principles. Section 3.2 describes the air defense testbed used to implement the experimental designs, as well as general information about the air defense task. Section 3.3 provides an overview of the theory of evidence introduced by Shafer (1976) that we used in the Phase II experiments to expand the Phase I screening condition so that it could deal with representative real-world conditions regarding both lack of information and conflicting information. Sections 3.4 and 3.5 describe the procedures and results for Experiments 1 and 2, respectively.

3.1 Basic Principles and a General Experimental Design

We begin this section by identifying those basic information-processing principles that the Phase I research suggests that people should implement to be good information processors. Our concern is only the high-workload condition because that is the condition under which adhering (or not) to these principles will have the biggest impact and, consequently, the effects of alternative human-machine allocation schemes are of most concern. We did, however, also use a low-workload condition in this first experiment. The focus through much of the discussion below is on the air defense problem being addressed in the Phase II experimentation, although the basic principles are, of course, general ones.

The basic principles are as follows:

(1) Decision makers should focus their attention on only those cases (in our case, targets) that require their attention. Other cases should be delegated to others (people or machine) that can solve them.

(2) Given that people are focusing on the cases needing their attention, they need to use an appropriate information-processing strategy to solve these cases. "Appropriate strategy" is a function of accuracy and speed.

These principles were supported in the Phase I study. The screening condition made Ps primarily attend to the unresolved targets and, to a lesser extent, more heavily rely on the "extra cue."

The implications of these principles were as follows for the Phase II experimentation when the person was working with the machine:

- When the machine's belief values are accurate, i.e., something has not happened that would cause them to be discredited:

- let the machine identify all targets it can (i.e., those that have passed specified thresholds on diagnosticity, uncertainty, and conflict);

- the person should use the appropriate strategy, such as examining cues in order of their value, to identify other targets (i.e., unknowns).

It was hypothesized that this implication would result in the following finding under the condition of high workload in our experiment:

- Any capability that permitted screening would improve performance over that which does not because it would help persons attend to the appropriate targets (i.e., unknowns). This "screening" capability could be in the form of both:

  - internal rules that let the machine identify targets that it can; and

  - an allocation menu that permits persons to create rules on-line that the machine can use to identify cases.

Two experiments were conducted to test this "implication" of the basic principles. The first experiment tested the relative effectiveness of three human-machine interface conditions (manual, override, and screening) under two levels of workload (low, high). Table 3-1 presents the design for the first experiment.

Table 3-1

Design for the First Phase II Experiment

		Workload Factor	
		Low	High
Human-Machine Interface Factor	Manual		
	Override		
	Screening		

The conditions were defined as follows:

(1) The human-machine interface factor (a "completely automated" condition was assumed as a baseline):

- *Manual* - Participants (Ps) identify all targets.

- *Override* - The machine identifies all targets; Ps change only those identifications they want to change.

- *Screening* - The machine firmly identifies all targets that have high levels of certainty; the machine prompts Ps as to which targets it can not identify (called

"unknowns") thereby forcing Ps to identify these unresolved (i.e., unidentified) targets; and the machine identifies "intermediate" targets, which are between firmly identified targets and unknowns, along with the reason for this judgment (e.g., not enough information or conflicting cues).

(2) The workload factor:

- *Low* - A target came on the screen every 12 seconds.
- *High* - A target came on the screen every four seconds.

Every participant performed the target identification task for all six cells of the 3 (human-machine interface) × 2 (workload) conditions, resulting in a within-participant design. This design replicates the Phase I experiment, but now using a significantly more representative air defense task (and target simulation) with actual air defense operators. Consistent with the Phase I results and "basic principles" above, we predicted:

(1) a significant main effect for workload, and

(2) a significant workload × human-machine interface interaction. Under low workload, all three interfaces would produce equal (and high levels) of performance. Under high workload, only the "screening" system would be able to maintain performance at the level achieved under low workload.

The second experiment tested the relative effectiveness of five human-machine interface conditions under high workload only. The experimental design is shown in Table 3-2.

Table 3-2

Design for the Second Phase II Experiment

		Allocation (i.e., Rule-Creation) Capability	
		No Allocation Capability	Allocation Capability
Human-Machine Interface Factor	Completely Manual	(Data inputted from 1st experiment)	
	Override		
	Screening		

Note: Assuming that "completely automated" is still a baseline condition.

Each participant ran all conditions, except the "manual:no allocation" cell of the design; the data for this cell were inputted from the first experiment to complete the design and analysis because time

constraints prevented us from running this cell. The conditions were implemented as follows:

(1) The human-machine interface factor (a "completely automated" condition was run as a baseline):

- *Manual* - Ps identify all targets as in the first experiment.

- *Override* - Machine identifies all targets and Ps change only those identifications they want to change, just as in the first experiment.

- *Screening* - Machine firmly identifies "certain targets" and human review is not needed; the machine passes off targets it can not identify (i.e., unknowns) to Ps for identification; and the machine highlights targets for which it made identification but human review suggested. (Note: In the first experiment, targets could leave the display as "unknown" if the P did not identify them, thereby becoming an incorrect identification. The screening condition was changed in the second experiment so that the machine always made its "best guess" identification for "unknowns" not identified by the P; consequently, no targets left the display as "unknown.")

(2) The allocation menu factor:

- *None* - Ps have no on-line mechanism for creating rules to be used by the machine to identify targets.

- *Allocation Capability* - Ps have on-line creation mechanism for creating rules to be used by the machine to identify targets.

Table 3-3 presents the hypothesized rank order of performance in high workload for the (Experiment 2) conditions described above; Figure 3-1 presents the same information pictorially.

Table 3-3

Hypothesized Rank Order of Performance (in High Workload) for Different Conditions

(Note: The lower the number, the better the performance.)

		Allocation Menu Factor	
		No Allocation Capability	Allocation Capability
Human-Machine	Completely Manual	4	3
Interface	Override	4	3
Factor	Screening	2	1

These predictions are consistent with the following principles:

(1) Performance in the override condition will not be much better than that achieved in the completely manual condition, and considerably less than that achieved in the screening condition, because it does not focus attention on the targets requiring it.

(2) Having an allocation (i.e., on-line rule-creation) capability will significantly improve performance for all three human-machine interface conditions.

(3) "Manual and override *with allocation*" will still not result in performance as good as that achieved with "screening without allocation" because participants are still not being directed to focus on certain targets requiring their attention.

(4) "Screening with allocation" will result in the best performance because it gives Ps the ability to create rules to identify certain classes of targets that they had to previously attend to because the machine did not have rules in its knowledge base for identifying these classes of targets. (Note: In the experiments, we used messages from Headquarters to alert Ps to classes of "unknown" targets for which rules could be created on-line.)

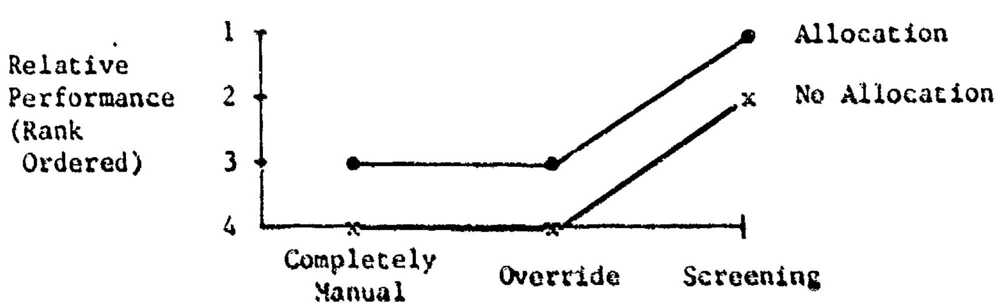


Figure 3-1. Hypothesized performance levels in high workload for the different conditions.

Specific details of the procedures for implementing each of the above two experimental designs, and the results for each experiment, are presented in Sections 3.4 and 3.5, respectively, of this report. We now turn to describing the air defense task and testbed developed for implementing the experiments.

3.2 The Air Defense Task and Experimental Testbed

The air defense environment is characterized by:

- high stakes, with expectations of heavy losses in short periods of time;

- potentially heavy peak loads, where the skies can be saturated with enemy aircraft;

- sophisticated, high-performance threat weapon systems, resulting in minimum reaction time;

- highly critical vital assets requiring friendly air defense protection.

The specific problem selected as the focus of the present research is identification of aircraft as friend or foe. Air defense systems exist in the field that run the gamut from virtually all identification (ID) decisions left in the hands of the user (e.g., IHAWK) to virtually all ID decisions made by the computer (e.g., PATRIOT). Current plans call for modifying the IHAWK system to improve its capabilities in the direction of PATRIOT, and there is a broad spectrum of intermediate capability options that must be considered.

During centralized operations, higher headquarters and adjacent units provide the dominant ID cues. At the fire unit level, targets are designated for engagement from battalion command centers, and the fire unit focuses on engagement rather than ID decisions. However, in wartime it is fully expected that a majority of combat engagements will be made under decentralized authority due to the inability of higher echelons to detect aircraft attacking at low altitudes. Additionally, it is expected that communications will be interrupted frequently, thus requiring fire units to operate autonomously.

In the IHAWK system, particularly during autonomous operations, the Tactical Control Officer (TCO) and Tactical Control Assistant (TCA) are responsible for identification decisions. Sources of information are varied:

- Identification Friend or Foe (IFF) equipment at the Platoon Command Post (PCP) (manually initiated by the TCA). This is done using a manual interrogation switch that is coded to receive specific responses from transponders in friendly aircraft. These codes are changed frequently to avoid exploitation by the enemy.

- Correlation with flight plans and safe-passage corridors--for example, the commander might establish a schedule which prohibits any firing at aircraft on certain headings during specified time periods.

- Aircraft actions (dropping chaff, use of other Electronic Counter Measures, or ECM, attacking friendly troops.

- Information passed from higher or adjacent units (if available).

- Pop-up criteria, which are designated parameters such as speed, altitude, and bearing that must be observed by friendly aircraft.

Often these identification cues are missing or can be conflicting, and the friend or foe decision is a difficult one. Following prescribed rules of engagement, the TCO/TCA typically will use electronic means to make a hostile identification based upon the above criteria. For example, if an aircraft is not responding to the prescribed IFF, is outside of a safe-passage corridor, and is closing at a speed in excess of a prescribed rate, it might be declared hostile. However, if it is not responding to IFF, is *in* the safe-passage corridor, and exceeds the speed criterion, the identification is less clearcut. Subjective judgment, based on experience, will be used to combine the ID cues and reach a decision.

Similarly, in the PATRIOT system, input data come from the sources above, except that a different IFF interrogator is used. Currently, cue conflict resolution depends upon the level of automation selected, the relative importance and reliability of the various data sources, and the decision-making style of the officer running the engagement. At the highest capability level, the automated system combines cues using a predetermined weighting algorithm and makes the ID determination. For example, IFF response, speed, altitude, and passage over restricted areas can be assigned weights that reflect their relative importance. Each aircraft is "scored" on each factor by the automated system, and based upon the mathematical combination of scores and weights, the aircraft is designated as friendly, hostile, or unknown. The TCO can change weights in the algorithm or can override automated decisions, but is not likely to do so in most cases. In this mode of operation, challenges using the IFF are initiated without TCO/TCA intervention.

A potential improvement to the IHAWK system would be to bring the capability level for ID closer to that of PATRIOT. More will be done by the automated portion of the system as far as challenging, analyzing data, and determining friend or foe identification. Yet, experience with the fully automated PATRIOT system has shown that this may not be the optimal configuration for the system. Some problems include excessive challenging to aircraft causing transponder damage, boredom of operators when in the fully automated mode, and lack of confidence in the fully automated system. Some have hypothesized that under combat conditions, TCO/TCAs will not even use electronic IFF challenges due to increased vulnerability to enemy exploitation of the signal. For these and other reasons, it may be the case that an intermediate level of automation will produce better results.

It is within this context that the research team designed the testbed for implementing the alternative conditions of the proposed experiment. We will now turn to discuss the testbed. When reading this description, one should keep in mind that certain characteristics of the "real air defense problem" (e.g., exactly when and how information is presented) were modified in order to obtain the experimental control necessary to evaluate Ps' information-processing strategies in alternative conditions. However, we tried to be as representative of the air defense environment as possible.

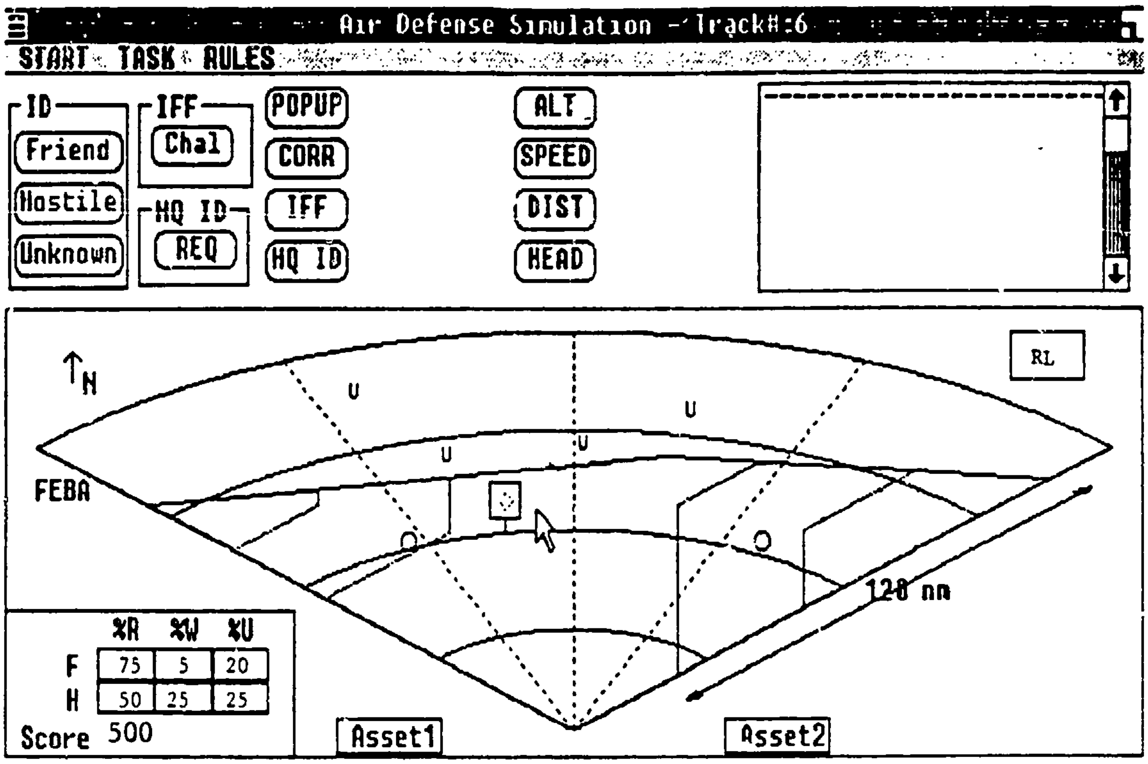
The "testbed" hardware was an IBM-AT personal computer with a color display monitor. The input devices were a three-key mouse and keyboard, although the Ps used only the mouse. The testbed had a 20-megabyte hard-disc for storing the air defense simulation and all software for presenting the experimental materials to the Ps for each condition and for collecting the Ps' responses.

The instructions to the participants can be found in Appendix A. In brief, they were told that they were the tactical control officer (TCO) of a proposed air defense system, and that they must decide which of the aircraft approaching them were friends and which were hostiles. They were shown a "radar" picture of the sector they were responsible for defending, and--on another portion of the screen--a variety of information that, in addition to the radar display, would help them make identification decisions. Figure 3-2 shows what the radar screen looked like. Conceptually, the P as the air defense operator, is located at the bottom, where the two straight lines come together. The whole pie-slice-shaped area is the area of responsibility of the unit. The P is protecting two friendly assets. In addition, there are two safe-passage corridors for the movement of friendly aircraft. These corridors were outlined in blue on the display.

The U-shaped figures on the screen shown in Figure 3-2 represent the location of "unknown" aircraft. These aircraft have not been identified as either friends or hostiles. Aircraft could fly at different speeds and altitudes depending on whether they were bombers, fighters, or helicopters. Aircraft could appear at the extreme top or sides of the screen or they could "pop-up" within the sector if they were flying at an altitude below radar detection. If they popped-up, they could do so "close" to the P or along the forward edge of the battle area (FEBA). Aircraft could be outside or inside a safe-passage corridor. Aircraft that were close to the edge of the corridor were classified as "inside" if they traced the edge of the corridor. If inside a corridor, they could be flying within the altitude and speed parameters set for the corridor or not. The acceptable corridor altitude ranged from 2,000 to 10,000 feet; the acceptable corridor speed was 700 knots or below. In all cases, the aircraft moved either toward the P or the sides of the sector; never toward its top. Ps were told that both friendly and hostile aircraft had been appearing for the first time well within the corridor and that, unfortunately, there had been heavy enemy jamming in the area of the corridor entrances and it had been impossible to use the corridor entry point as a reliable ID criterion.

Ps were told that, in order to perform well, they had to correctly identify as many aircraft (i.e., friend or hostile) as possible before the aircraft went off the radar screen. Aircraft went off the screen when they had reached the sides of the sector or when they were 40 km from the P's position, which is the closest range ring in Figure 3-2. Ps were told that they were responsible for, and scored on, all targets within the FEBA, even though their engagement capability extended only to the 2nd range ring.

Figure 3-2 shows the symbols for different types of targets. A black "U" represented an "unknown;" a circle represented a "friend;" a diamond represented a "hostile." A "hexagon outline" represented a target that had



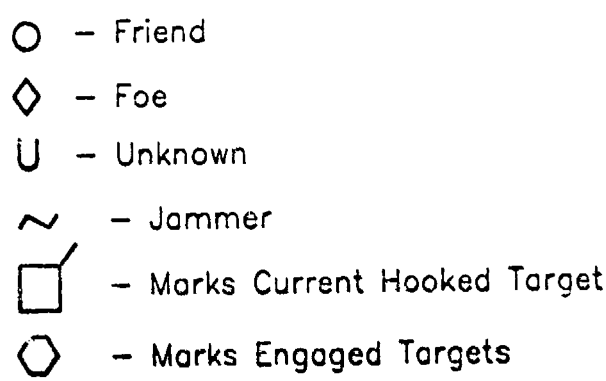


Figure 3-2. The radar screen and basic symbols.

been engaged by the system. Jammers were indicated by the symbol shown in Figure 3-2.

Points were used to facilitate Ps' motivation. Specifically, Ps received 5 points for each correctly identified target, either friend or hostile. They received no points if they identified the target incorrectly or if they left the target as "unknown." It was emphasized that their goal was to maximize their point total. To help them, Ps were given feedback every minute. The feedback was presented in a box in the lower left-hand portion of the screen. During an attack phase, it told the Ps what cumulative proportion of friends and hostiles they identified correctly, incorrectly, or left as unknown when those aircraft left the screen. If a high proportion of "true" friendly aircraft were identified incorrectly, the P was identifying too many friends as hostiles. Consequently, the P needed to examine aircraft on the screen that were identified as hostile more carefully because some of them were friends. In contrast, if a high proportion of "true" hostile aircraft were identified incorrectly, the Ps needed to examine aircraft on the screen that were identified as friends because some of them were probably hostiles.

At the end of an attack phase, the feedback box added the proportion of friends and hostiles *within the FEBA* identified correctly, incorrectly, or left unknown to help the P determine how well he did for that phase. Attack phases varied in the amount of time they took, depending on the workload condition.

In order to identify a target or obtain more information about it, Ps "hooked" the target by (1) using the mouse to guide a cursor to the target, and (2) pressing the mouse's left-hand button. Only one target could be hooked at a time. "Hooked targets" were represented by a square on the radar screen; their track #s appeared at the top of the radar scope, as shown in Figure 3-3.

Ps could use the other "buttons" on the top-half of the screen to obtain more information about a hooked target. In particular, they could find out whether the target "popped-up" (POPUP button); whether it was "in" or "out" of the corridor (CORR button); or its speed (SPEED button), distance (DIST button), altitude (ALT button), and heading (HEAD button).

There are two buttons labeled "IFF CHAL" and "HQ ID REQ" in the upper portion of the main display. Ps could gather *new* information about an aircraft by pressing these buttons. "IFF CHAL" stands for "IFF Challenge," which is sending an electronic interrogation signal to which friendly aircraft can respond automatically unless their equipment has malfunctioned, they do not have an IFF transponder, or their codes are set improperly. Although typical air defense systems simultaneously challenge multiple targets, an IFF challenge could be placed only against the hooked target in the testbed in order for us to examine the Ps' information-processing (and search) strategies. The "HQ ID REQ" button was used to contact higher headquarters (HQ) to ask them for the identification of the hooked target. The answer to the P's request appeared to the right of the button. If the P received "unknown" in response, this meant that headquarters did not know the identification of the aircraft. Figure 3-3 shows

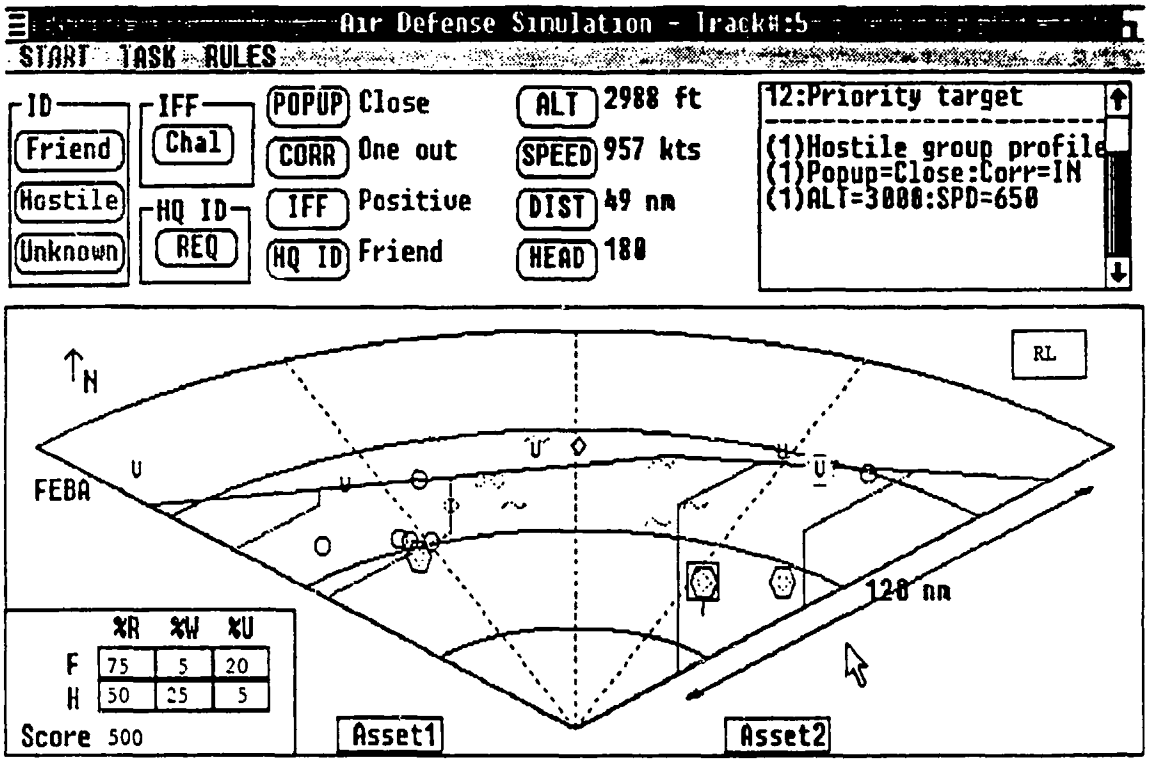


Figure 3-3. The entire display screen.

all the information for the "hooked aircraft" in the right-hand safe-passage corridor.

The Ps were told that, in general, there were more hostiles than friendlies, and that some of the information (called "cues") were better discriminators than others. Ps were shown the table below, which indicates which cues were the strongest indicators for each type of target, and which had a lot of uncertainty associated with them in the testbed's simulation. Although the diagnosticity of the cues in our simulation match those in an actual air defense environment moderately well, distinct differences (e.g., that for "HQ ID Hostile" or "Corridor Two Out") did develop in the simulation because of the effort to create "test cases," that is, targets that would be difficult to identify. We showed Ps the diagnosticity table to minimize the extent to which these differences affected Ps' performance.

Cues Indicating FRIEND		Cues Indicating FOE		Non-Definitive
Value	Strength	Value	Strength	
IFF Positive	Very Strong	Corridor OUT	Strong	HQ ID Unknown
HQ ID Friend	Very Strong	HQ ID Hostile	Strong	Popup (all values)
Corridor IN	Strong	Jammer	Strong	Non-Jammer
Corridor One Out	Moderate	IFF No Response	Moderate	Corridor Two Out

As can be seen from the above table, performing an "IFF Challenge" and a "HQ ID Request" were by far the most diagnostic pieces of information. Two actions were taken to prevent Ps from mechanically hooking a target and pressing the buttons for those two pieces of information in order to increase the judgmental aspects of the problem. First, penalties were attached to "IFF Challenge" and "HQ ID Request." In particular, a point penalty was attached to the former and a time penalty to the latter.

Performing an IFF challenge in the "real world" opens an air defender to exploitation (i.e., attack) by enemy aircraft. To represent this situation, Ps were (randomly) exploited 10% of the times they issued an "IFF Challenge." If exploited, they lost 10 points and were notified in the lower left-hand corner of the display. Requesting an "HQ ID" takes time to perform in an actual air defense environment. To represent this, it took 4 seconds for Ps to get a response to their requests. During this time they could get information about the hooked target, but they could not hook another target. If the Ps performed an "IFF CHAL" or requested an HQ ID for a hooked target, they could recall the information from their database at a later time without incurring a point or time penalty. This was done by rehooking the target and hitting the "IFF" and "HQ" buttons located beneath the "SPEED" and "CORR" buttons.

The second action taken was to make available information particularly salient for certain targets so that there would be no need to press the "IFF CHAL" and "HQ ID REQ" buttons. This was accomplished by sending Ps messages from headquarters about certain targets. This information appeared in the MESSAGE box. For example, a message might have looked as follows:

(1) Hostile Group Profile
(1) Popup=No:Corr=Two out
(1) Alt=11,000:Speed=1200

The number in the parentheses indicated that this was Message #1. "Hostile Group Profile" meant that the message described a group of targets that was known to be hostile. "Popup=No:Corr=Two out" told Ps that the hostiles did not pop-up and that, although they could be within the lateral boundaries of both corridors, their speed *and* altitude did *not* match the parameters of the safe-passage corridors. In particular, the hostiles had an altitude of approximately 11,000 feet and a speed of approximately 1200 knots. No friendly aircraft were at that altitude and speed specified in the messages. The message was in effect for no more than ten minutes, depending upon how long it took the hostile group to leave the radar screen. (Note: After those ten minutes, friendly aircraft might appear at the altitudes and speeds identified in the message.)

We used a secondary task to obtain an objective measure of workload. Specifically, while Ps were performing the target identification task, they had to acknowledge orders from higher headquarters. These orders were represented by a light in the upper right-hand corner of the radar display. It is labeled "RL" in Figures 3-2 and 3-3. This light went on throughout the session based on a Poisson distribution. It stayed on for 3 seconds. In an effort to ensure that Ps responded to the secondary task, (a) Ps were told to acknowledge the RL as fast as possible, and (b) if they failed to respond within 3 seconds, they lost 2 points. We obtained accuracy measures by asking Ps to press the middle button on the mouse when the light was red, and the right-hand button on the mouse when the light was green. The buttons were color-coded. Ps obtained 1 point every time they responded correctly within the time limit; they lost 1 point if they responded incorrectly.

The Ps' total "running" score appeared at the bottom of the feedback box in the lower left-hand corner of the display. The total score was a function of: (1) the number of aircraft identified correctly; (2) the number of points lost through the exploitation of IFF challenges; and (3) the number of points gained or lost through acknowledgments of the response light. Ps were told that the *worst possible* score they could obtain at the end of an attack phase was -400 points and that the *best possible* score was +1500 points in an effort to help them evaluate their performances.

We now turn to a description of Shafer's theory of evidence, which was used and expanded upon by the research team to guide the machine's inference process. The material below is, in large part, an abridged version of a much more detailed discussion found in Cohen, Watson, and Barrett (1985).

3.3 Shafer's Theory of Evidence

In the theory of belief functions introduced by Shafer (1976), Bayesian probabilities are replaced by a concept of evidential support. The contrast, according to Shafer (1981; Shafer and Tversky, 1983) is be-

tween the chance that a hypothesis is true, on the one hand, and the chance that the evidence *means* (or proves) that the hypothesis is true, on the other. Thus, we shift focus from truth of a hypothesis (in Bayesian Theory) to the evaluation of an evidential argument (in Shafer's Theory). By stressing the link between evidence and hypothesis, Shafer's system (a) is able to provide an explicit measure of quality of evidence or ignorance (i.e., the chance that the evidence is not linked to the hypothesis by a valid argument); (b) is less prone to require a degree of definiteness in inputs that exceeds the knowledge of the expert, and (c) permits segmentation of reasoning into analyses that depend on independent bodies of evidence. We will find that each of these properties can contribute significantly to the representation of uncertainty.

In Shafer's system, the support for a hypothesis and for its complement need not add to unity. For example, if a witness with poor eyesight reports the presence of an enemy antiaircraft installation at a specific location, there is a certain probability that his eyesight was adequate on the relevant occasion and a certain probability that it was not, hence, that the evidence is *irrelevant*. In the first case, the evidence proves the artillery is there. In neither case could the evidence prove the artillery is not there.

To the extent that the sum of support for a hypothesis and its complement falls short of unity, there is "uncommitted" support, that is, the argument based on the present evidence is unreliable. Evidential support for a hypothesis is a lower bound on the probability of its being true, since the hypothesis *could* be true even though our evidence fails to demonstrate it. The upper bound is given by supposing that all present evidence that is *consistent* with the truth of the hypothesis were in fact to prove it. The interval between lower and upper bounds--that is, the range of permissible belief--thus reflects the unreliability of current arguments. This concept is closely related to *completeness* of evidence, since the more unreliable an argument is, the more changeable the resulting beliefs are as new evidence (with associated arguments) is discovered. These concepts are not directly captured by Bayesian probabilities.

In Shafer's calculus, support $m(\cdot)$ is allocated not to hypotheses, but to *sets* of hypotheses. As with probability, however, the total support across these subsets will sum to 1, and each support $m(\cdot)$ will be between 0 and 1. It is natural, then, to say that $m(\cdot)$ gives the probability that what the evidence *means* is that the truth lies somewhere in the indicated subset.

Suppose, for example, that we have three hypotheses of interest: H_1, H_2, and H_3. We feel the evidence *either* means that H_3 is true, *or* that (H_1 or H_3) is true, *or* that it is not telling us anything (i.e., (H_1 or H_2 or H_3) is true), and that the weight of evidence is just as strong with each possibility. In that case, $m(H_3) = m((H_1 \text{ or } H_3)) = m((H_1 \text{ or } H_2 \text{ or } H_3)) = 1/3$. In a Bayesian analysis, arbitrary decisions would have to be made about allocating probability *within* these subsets, requiring judgments that are unsupported by the evidence.

This same device, of allocating support to subsets of hypotheses, enables us to represent the reliability of probability assessments. Suppose, for example, that a certain type of seismic reading has been associated with seismic activity 70% of the time and with nuclear tests 30% of the time based on past frequency data. If we are confident that seismic data now being analyzed are representative of this set, we may have m(earthquake) = .7 and m(nuclear test) = .3. But if there is reason to doubt the relevance of the frequency data to the present problem (e.g., because the frequency data come from U.S. tests and seismic data from other geological regions), we may *discount* this support function by allocating some percentage of support to the universal set. For example, with a discount rate of 30%, we get m(earthquake) = .49, m(nuclear test) = .21, and m({earthquake, nuclear test}) = .30. The latter reflects the chance that the frequency data are irrelevant.

Similarly, within the context of the air defense experiment described above, the unreliability of various means of intelligence collection will reduce one's belief in HQ ID. For example, we might say that "HQ ID = Friend" has a level of unreliability such that $m(H_1$:target is friend) = .8, $m(H_2$:target is foe) = 0, $m(H_3$:target is uncertain, i.e., either friend or foe) = .2. Moreover, it is possible that "HQ ID = Foe" is more unreliable than "HQ ID = Friend." Consequently, one might obtain the following degree of support values for "HQ ID = Foe": $m(H_1$:target is friend) = 0, $m(H_2$:target is foe) = .60, and $m(H_3$:target is either friend or foe) = .40. In developing the belief functions for the testbed, we took the direct approach of independently asking two substantive domain experts the "Degree of Support for Friend," the "Degree of Support for Hostile," and the "Degree of Support for Either, i.e., Don't Know," for each level on each cue. We then met jointly with both experts to resolve differences.

Shafer's belief function Bel(·) summarizes the implications of the m(·) for a given subset of hypotheses. Bel(A) is defined as the total support for all subsets of hypotheses contained within A; in other words, Bel(A) is the probability that the evidence *implies* that the truth is in A. The plausibility function Pl(·) is the total support for all subsets which overlap with a given subset. Thus, Pl(A) equals $1\text{-Bel}(\bar{A})$; i.e., the probability that the evidence does not imply the truth to be in not-A. In the example above, we get:

$$\text{Bel}(H_3) = m(H_3) = 1/3;$$

$$\text{Pl}(H_3) = 1\text{-Bel}(\{H_1 \text{ or } H_2\}) = 1;$$

$$\text{Bel}(\{H_1 \text{ or } H_3\}) = m(H_3) + m(\{H_1 \text{ or } H_3\}) = 2/3;$$

$$\text{Pl}(\{H_1 \text{ or } H_3\}) = 1\text{-Bel}(\{H_2\}) = 1.$$

Thus far, we have focused on the representation of uncertainty in Shafer's system. For it to be a useful calculus, we need a procedure for inferring degrees of belief in hypotheses in the light of more than one piece of evidence. This is accomplished in Shafer's theory by Dempster's

rule. The essential intuition is simply that the "meaning" of the combination of two pieces of evidence is the intersection, or common element, of the two subsets constituting their separate meanings. For example, if evidence E_1 proves $\{H_1$ or $H_2\}$, and evidence E_2 proves $\{H_2$ or $H_3\}$, then the combination $E_1 + E_2$ proves H_2. Since the two pieces of evidence are assumed to be independent, the probability of any given combination of meanings is the product of their separate probabilities.

Let X be a set of hypotheses $H_1, H_2, \ldots, H_n$, and write 2^X for the power set of X, that is, the set of all subsets of X. Thus, a member of 2^X will be a subset of hypotheses, such as $\{H_2, H_5, H_7\}$, H_3, or $\{H_1, H_2, H_3, H_4\}$, etc. Then if $m_1(A)$ is the support given to A by one piece of evidence, and $m_2(A)$ is the support given by a second piece of evidence, Dempster's rule is that the support that should be given to A by the two pieces of evidence is:

$$m_{12}(A) = \frac{\sum_{A_1 \cap A_2 = A} m_1(A_1)m_2(A_2)}{1 - \sum_{A_1 \cap A_2 = \phi} m_1(A_1)m_2(A_2)}.$$

The numerator here is the sum of the products of support for all pairs of subsets A_1, A_2 whose intersection is precisely A. The denominator is a normalizing factor which ensures that $m_{12}(\cdot)$ sums to 1, by eliminating support for impossible combinations.

The m("Degree of Support for Either, i.e., Don't Know") represents the level of uncertainty in the evidence, and this value can range from 0 to 1.0 such that the sum of the evidential values equals 1.0. In addition to uncertainty, there is conflict when different pieces of evidence occur supporting different hypotheses (e.g., friend and foe). Human experts typically use conflict as a symptom of the existence of problems in the analysis, and thus as a prompt for action (such as reexamining the credibility of sources, reconsidering basic assumptions of the analysis, or searching for new information). The simulation for the experiment was designed such that significant conflict occurred for certain types of targets. It was neither clear exactly what information-processing strategy Ps would use to resolve conflicts, nor how the various conditions in the experiment would affect their strategy. Nor was it clear that Ps would use the same strategy to resolve "uncertain" versus "conflict" cases. Shafer's theory of evidence gives one the framework of differentiating between these two types of cases and, more globally, the means for determining whether Ps' information-processing strategy (and conclusions) match those proposed by the theory.

3.4 Experiment 1

This section of the report describes the method and results for the first experiment conducted at Fort Bliss, Texas.

3.4.1 Method. The method section for Experiment 1 is composed of the following subsections: the experimental design, the participants, how the factors in the design were operationalized, a description of the air defense simulation, the procedures used when conducting the experiment, and the dependent measures used to assess the effects of the different experimental conditions. Each subsection is considered in turn.

3.4.1.1 *Experimental Design*. A repeated measures, 2 workload (low and high) × 3 human-machine interface (manual, override, and screening) factorial design was used in Experiment 1. A completely automated condition where the system performed the target identification task without access to IFF, HQ ID, or message data served as a baseline condition. This design replicated the Phase I experiment, but now using a significantly more representative air defense task (and target simulation) with actual air defense operators.

3.4.1.2 *Participants*. Fourteen U.S. Army air defense operators participated in the experiment between 19 May-29 May 1987. All participants were either first or second lieutenants who had completed the Basic Course and who had some experience with either the PATRIOT or HAWK air defense system.

3.4.1.3 *Operationalizing Design Factors*. Workload levels were operationalized by manipulating the time in between which targets appeared on the radar screen. Specifically, a target appeared every 12 seconds in the low-workload condition and every 4 seconds in the high-workload condition. This 3:1 ratio was quite similar to the 2.75:1 ratio used in the Phase I experiment, and easy to operationalize on the testbed.

The manual, override, and screening conditions were operationalized as follows. In the *manual* condition, Ps were told that the system would keep track of all information about the targets but that *they had to perform all target identifications*. The concept of "conflicting information" was presented in the form of the example of an aircraft that is jamming--which suggests it is hostile--but giving a positive IFF response--which suggests it is friendly. Ps were reminded about how to obtain information, and about the point and time penalties for "IFF CHAL" and "HQ ID REQ," respectively. In sum, the manual condition in Experiment 1 was operationalized basically like it was in the Phase I study, except for changes representative of the more sophisticated defense task and simulation.

In the *override* condition, Ps were told that, in addition to keeping track of all the information about the targets, the system would also make an initial identification of *all* aircraft based on: (a) whether and where it popped up; (b) whether it was in the corridor or not; (c) whether its speed and altitude met the corridor parameters if it was in the corridor; and (d) whether or not it was a jammer. Aircraft initially identified as a friend were represented as black circles. Aircraft, exept jammers, initially identified as hostile were represented as black diamonds. All jammers were initially identified as hostile by using the black jammer symbol. It is important to note that the system did not have access to messages from headquarters or the results of an HQ ID or IFF challenge when it made the initial identification. Consequently, the override condition was

operationalized similarly to how it was operationalized in the Phase I study where the system used all the information except the "extra cue," which had to be searched for and processed by the human operator. One important difference between the override and screening conditions in Experiment 1 and the Phase I study, however, was that the responses to the IFF challenge were included in the system's identification algorithm in the former in an effort to help maintain task representativeness with actual air defenders.

Ps, when in the override condition, were told about "conflicting information" and reminded about penalties just as in the manual condition. Ps also were told that unless they changed the system's identification, it would represent their identification when the aircraft went off the screen or, if the aircraft was within the FEBA, when the attack session ended. Changes in the identifications made by the Ps were color-coded. In particular, a blue circle represented an aircraft Ps identified as friend; a red diamond represented an aircraft that Ps identified as hostile. Ps were also told that if they changed an identification to "unknown," because they wanted to identify another aircraft before deciding, it would be represented as a green U, and reminded that they would not get any points for aircraft identified as "unknown." If Ps identified a jammer, the jammer symbol was color-coded blue, red, or green depending on whether it was identified as friend, hostile, or unknown, respectively.

Finally, when in the override condition, Ps were told that in certain cases where aircraft moved into the corridor or out of the corridor, the machine might change an identification from hostile to friend, or vice versa. This could be either a good or bad action. It could be a good action if, for example, the P made an identification before an aircraft entered a safe-passage corridor and, after the aircraft entered the corridor, the machine took this information into account. It could be a bad decision if the machine changed an identification the P made on the basis of an HQ ID. Ps were told to remember that the machine did not have access to the results of an HQ ID. Consequently, if they identified a target as a friend on the basis of an HQ ID and the aircraft left the corridor for whatever reason, the machine could incorrectly change their identification. Whenever the machine changed an ID, a message appeared in the message box informing the P of this change. Ps could 'click' on this message with the mouse to hook this target and then again identify the target.

When in the *screening* condition, Ps were again told that the system kept track of all the information about all the targets, and that it also made an initial identification of *all* aircraft based on: (a) whether and where it popped-up; (b) whether it was in the corridor or not; (c) whether its speed and altitude met the corridor parameters if it was in the corridor; and (d) whether or not it was a jammer. Again, it was noted that the system did not have access to messages from headquarters or an HQ ID, and it could not initiate IFF challenges. On the basis of information it did have, the system used a blue circle to identify aircraft that clearly appeared to be friendly; it used a red diamond (or jammer symbol) to identify aircraft that clearly appeared to be hostile. "Firm identifications" were targets that had degrees of belief in Shaferian terms (see Section [illegible]

If the system were less certain of its identification, it used a black circle to identify "questionable friends" and a black diamond to identify "questionable hostiles." By "questionables" we meant there was not enough information to firmly ID, but the evidence was more in favor of one type or another ($.6 \leq$ degree of belief $< .8$). The color *black* meant there was no conflicting evidence. When there was conflicting information, the system used the color *purple*. A purple circle represented an aircraft that was a "questionable friend" because of conflicting information. A purple diamond (or jammer symbol) represented an aircraft that was a "questionable hostile" because of conflicting information. (Note: The colors black and purple were used to suggest the reason why the system classified different targets as questionable. All statistical analyses of performance were, however, performed on the entire subset of N=57 questionable targets because we wanted to know whether the screening condition improved performance for this class of targets, irrespective of the reason for the classification.)

If, either because there was not enough information or the information was conflicting, the system was unable to identify the aircraft on the basis of its initial information (i.e., degree of belief $< .6$), the system in the screening condition classified the aircraft as an "unknown" (a black U). Often, however, the system indicated a "highest priority unknown." This was the target that, in the system's opinion, was the most important to ID next. The priority rating was based on the amount of uncertainty, the amount of conflict, and the aircraft's "time to nearest friendly asset." This unknown (U) had a purple, solid circle around it. In addition, its identification number appeared at the top of the message box. Ps could hook the "highest priority unknown" by either (1) clicking on its identification number in the message box, which hooked it automatically, or (2) hooking it just like any other aircraft. It is important to note that aircraft could go off the screen classified as "unknown" if the P failed to identify them. This is different than in the Phase I study where Ps only had to identify threats. Consequently, in the Phase I study, it was not possible to ascertain whether an unknown was an aircraft that the P was unable (or did not have time) to identify (i.e., a true "unknown") or ar aircraft that the P knew was a friend, but did not bother to identify. Aircraft had to be explicitly identified by Ps (with or without the system) in all conditions in Experiment 1.

Finally the screening condition had all of the other capabilities available in the override condition.

3.4.1.4 *Simulation*. The basic components of the air defense simulation from the participants' perspective are described in considerable detail in Section 3.2 of the report; consequently, only those aspects that affected the ability to analyze performance under various experimental conditions are discussed here. Specifically, a 200-target simulation was developed for subsequent test and analysis. Twenty-five targets preceded these 200 targets under all conditions as a means of getting participants actively involved in the air defense task before any targets for whom participants' performance would be analyzed appeared on the screen. All 200 (performance) targets left the screen before the simulation ended in all conditions. The number of targets following the 200 (performance) targets

depended, however, on the workload condition because targets came on the screen faster in the high-workload condition, and yet, each target still took the same amount of time to traverse its path across the radar screen.

The 200 (performance) targets, hereafter referred to as the "performance simulation," were constructed of three sets (or classes) of targets on the basis of the data available to the system in the completely automated condition, that is, prior to IFF challenges, HQ ID requests, or messages. The three sets of targets directly match the three identification categories in the screening condition: "firm identification" (degree of belief $\geq .80$), "questionables" ($.6 \leq$ degree of belief $< .8$ regardless of conflict), and "unknowns" (degree of belief $< .6$). In particular, there were 85 "firm identification" targets, 57 "questionable" targets and 56 "unknowns." In the completely automated condition, the computer correctly identified 74 (or 87%) of the "firmly identified targets," 36 (or 64%) of the "questionables," and 27 (or 46%) of the "unknowns." The "unknowns" are most comparable to the test cases in the Phase I experiment, which had a $p = .50$ accuracy without the extra cue.

It should be noted here for its implications for future research that development of the performance simulation was a slow and time-consuming task composed of four principal activities. The first activity was to divide targets into the three target sets described above. Two DSC domain experts were interviewed to obtain degree-of-belief (DB) values indicating the extent to which each cue level indicated (or pointed to) friend, foe, or don't know. The domain experts were independently interviewed primarily by one knowledge engineer (others assisted in the knowledge engineering [illegible]s) over a number of sessions to obtain independent assessments of the beli[illegible]ues. Meetings were then held with both domain experts to resolve areas of disagreement. The belief values used to construct the performance simulation and used in the override and screening conditions are shown in Table 3-4.

Table 3-4

Degree of Belief Values for Individual Cue Levels
(Assuming Cue Independence)

CUE		DEGREE OF SUPPORT FOR "FRIEND"	DEGREE OF SUPPORT FOR "HOSTILE"	DEGREE OF SUPPORT FOR "DON'T KNOW"
IFF	RESPONSE	98	0	2
	NO RESPONSE	0	55	45
JAMMING	YES	0	95	5
	NO	0	0	100
POP-UP	NO	0	0	100
	CLOSE	0	80	20
	FEBA	0	60	40
OUT OF CORRIDOR		0	80	20
HEADING IN CORRIDOR	SPEED AND ALTITUDE CORRECT	90	0	10
	SPEED OR ALTITUDE WRONG	75	5	20
	SPEED AND ALTITUDE WRONG	50	20	30
NO ID	FRIEND	95	0	5
	HOSTILE	0	90	10
	UNKNOWN	0	0	100

Then, using the values for the different levels of pop-up, jamming, and corridor and the Dempster-Shafer algorithm described in Section 3.3 of the report (as implemented on a spreadsheet), we created the three target types (firmly identified, questionable, and unknown) described above.

The second principal activity was to determine the values for IFF challenge and HQ ID for targets with the same levels on the pop-up, jamming, and corridor cues so that the proportion of friends and hostiles in the simulation after IFF Challenge and HQ ID would approximate the degrees of belief for friend and hostile prior to this information. This was important in an effort to ensure the overall representativeness of the performance simulation to an actual air defense environment, as represented by the system's degree-of-belief values, and yet create targets with conflicting information for the experiment. For example, according to the (compromise) degree-of-belief values of the two participating domain experts and the Dempster-Shafer algorithm for combining these values into overall values, a target that, prior to an IFF response on an HQ ID, (a) pops-up close in the safe-passage corridor, (b) is not jamming, and (c) is at the correct speed and altitude should, in general, have an overall degree of belief (Friend) = .64, DB (Hostile) = .29, and DB (Don't Know) = .07.

Table 3-5 shows the table we used to construct the IFF challenge responses and HQ IDs for the 14 targets in the performance simulation so that 9 of them (i.e., 64%) were actually friends and the degrees of belief pointed to the correct action.

Table 3-5

The IFF Challenge and HQ ID Responses Constructed for One Type of Target

VARIATION	POP-UP	JAMMING	CORRIDOR	IFF CHALLENGE	HQ ID	DB (FR)	DB (FOE)	DB (?)	AMOUNT OF CONFLICT	# OF FRN	# OF FOES
START	YES, CLOSE	NO	YES, CORRECT	NO	NO	.64	.29	.07	.72	9	5
#1	"	"	"	RESPONSE	FRN	1.00	0	0	.80	1	0
#2	"	"	"	RESPONSE	UNK	.99	.01	0	.80	1	0
#3	"	"	"	RESPONSE	FOE	.91	.09	0	.98	1	0
#4	"	"	"	NO RESPONSE	FRN	.94	.05	.01	.91	5	0
#5	"	"	"	NO RESPONSE	UNK	.45	.50	.05	.82	2	2
#6	"	"	"	NO RESPONSE	FOE	.07	.92	.01	.89	0	3

Table 3-6 shows the cue diagnosticities--i.e., normalized likelihood ratios--for individual cue levels based on the entire 200 target performance simulation. Comparing Table 3-4 with Table 3-6 shows good comparisons in some cases (e.g., IFF response, out of corridor, heading in corridor when speed and altitude correct, etc.), but only moderate comparisons (at best) in other cases (e.g., HQ ID Hostile and Jamming).

Consequently, we concluded that we were only reasonably successful at maintaining the representativeness of the air defense environment. For this reason, participants were given qualitative information about the diagnosticity of individual cue levels while being trained to perform their basic job on the system regardless of the human-machine interface condition.

Table 3-6

Diagnosticity Values for Individual Cue Levels
(Based on all Targets in Performance Simulation)

CUE		DIAGNOSTICITY FOR "FRIEND"	DIAGNOSTICITY FOR "HOSTILE"
IFF	RESPONSE	1.00	0
	NO RESPONSE	0	.84
JAMMING	YES	.23	.77
	NO	.59	.31
POP-UP	NO	.65	.35
	CLOSE	.43	.57
	FEBA	.36	.64
OUT OF CORRIDOR		.13	.87
HEADING IN CORRIDOR	SPEED AND ALTITUDE CORRECT	.92	.08
	SPEED OR ALTITUDE WRONG	.76	.24
	SPEED AND ALTITUDE WRONG	.13	.87
HQ ID	FRIEND	1.00	0
	HOSTILE	.21	.79
	UNKNOWN	.24	.66

The third principal activity involved in developing the simulation was to create messages so that 28 of the 58 "unknown" targets and 28 of the 57 "questionable" targets could be perfectly identified without IFF challenges or requests for HQ ID. This was accomplished by having the messages focus on groups of targets that were in the safe-passage corridor, but that had particular values for pop-up, speed, and altitude. This action was taken in an attempt to make participants have a reason for using this information and not simply perform the identification task solely on the basis of visual information (i.e., being out of the corridor and/or jamming) or HQ ID and IFF challenges.

The fourth principal activity was developing the file of instructions for telling each target where to appear on the screen and how to traverse its path across it. A file of information, referred to as the master file, was created. The master file contained a target identification number, the target's cue value for IFF, HQ ID, Pop-Up, Jamming, and Corridor, and the target's true identification. The master file also contained the target type (fighter, bomber or helicopter), the initial location of the target in terms of distance from self and azimuth, the initial heading of the target in degrees north, and speed. No target was permitted to head in an upward direction, that is, all targets had an initial heading randomly chosen to be between 135 and 225 degrees. Generally, targets that did not pop-up were initially located beyond the FEBA and targets that popped-up close

were initially located between the FEBA and the second inner ring. Targets that popped-up near the FEBA and had corridor values of in, one out, or two out initially appeared in the top part of either of the two corridors. Targets that popped-up close and had corridor values of in, one out, or two out initially appeared in the second part of either of the safe-passage corridors. Targets that were not pop-ups but had corridor values of in, one out, or two out were initially located beyond the FEBA and moved into one of the two corridors. Targets were randomly assigned their characteristics based on the following guidelines.

Guidelines for "one out" targets that pop-up.

Altitude:	2000-3000	
Speed:	Bomber:	700-1000
	Fighter:	700-1500

Guidelines for "two out" targets that pop-up.

Altitude:	100-2000	
Speed:	Bomber:	700-1000
	Fighter:	700-1500

Guidelines for "in" corridor targets that pop-up.

Altitude:	2000-3000	
Speed:	Bomber:	400-500
	Fighter:	600-699
	Helicopter:	150-220

Guidelines for "one out" targets that do not pop-up.
Target may have either an invalid speed or an invalid altitude.

Invalid Speed

Altitude:	2000-10000	
Speed:	Bomber:	700-1000
	Fighter:	700-1500

Invalid Altitude

Altitude:	100-2000	
Speed:	Bomber:	400-500
	Fighter:	600-699
	Helicopter:	150-220

Guidelines for targets "in" the corridor that do not pop-up.

Altitude:	2000-10000	
Speed:	Bomber:	400-699
	Fighter:	600-699
	Helicopter:	150-220

Guidelines for targets that pop-up close and are "out" of the corridor.

Pop-Up in the Closest Friendly Half.

Altitude:	300-3000	
Speed:	Bomber:	400-500
	Fighter:	400-500
	Helicopter:	150-220

Guidelines for targets that pop-up near the FEBA and are "out" of the corridor.

Pop-up Near the FEBA.

Altitude:	300-3000	
Speed:	Bomber:	400-500
	Fighter:	600-700
	Helicopter:	150-220

The exceptions to these general guidelines were the test cases with messages. These targets were assigned random values for speed and altitude within a small range so that they could be identified based on the speed and altitude values appearing in the message. The initial location of each target was checked to ensure that its appearance on the screen was consistent with the targets assigned cue values. Targets were reassigned initial locations and/or speeds if, given their current characteristics, they could not travel off the scope within ten time steps.

The information in the master file was then used as input into a computer program that determined each target's path across the radar scope. For each sweep of the radar scope, the program produced a new location for each target based on its current heading and speed. The program also determined if the current cue values changed based on the target's new location. For example, the program determined which direction a "no pop-up" target with corridor value of in, one out, or two out would have head in order to move into the closest corridor. (This target would initially appear beyond the FEBA and would move into a corridor.) At each sweep of the radar scope, the program ensured that targets supposed to be in the corridor were indeed in the corridor and ensured that the target continued to trace the corridor, once in it. The program produced an output file containing each target's location at each of the ten sweeps of the radar scope. Every target in the simulation traveled off the radar scope by the tenth time step. The file also contained each target's current cue value at each time step, and the associated belief function with and without an IFF Challenge, because we did not know when or if a participant would perform an IFF Challenge on a target. All this information was stored in the target file. The air defense experiment program then read in the information concerning location, cue values, and belief functions to assess the target's identification and move each target across the radar scope.

The targets were displayed to the user in the order depicted in one of six order files. The three order files for the high-workload conditions contained 355 total targets: 25 were leading targets; 200 made up the performance simulation; and the rest were trailing targets. The number of trailing targets for the high-workload condition targets ensured that all 200 performance targets traveled off the radar scope before the simulation ended. The three low-workload order files contained 265 targets: 25 were leading targets; 200 made up the performance simulation; and the rest were trailing targets. Again, the number of trailing targets for the low-workload condition ensured that all 200 performance targets had left the radar scope before the simulation ended. The order of the leading targets, performance targets, and trailing targets was randomly determined for each of the six files. Then the six "order files" were randomized as part of

the experimental design, and the experimenters were provided with instructions as to the particular order file to use.

3.4.1.5 *Procedures*. Each of the fourteen participants took an entire work-day (approximately 8 hours not including a one-hour lunch break) to perform the air defense task for each of the six cells in the 2 Workload x 3 Human-Machine Interface design. Most of the time, two participants participated each day the experiment was conducted at the ARI Field Unit's offices at Fort Bliss. The participants were separated by partitions, and they worked independently on identical testbed systems. There were two experimenters, one for each participant.

The experimenters were graduate students in the Department of Psychology at New Mexico State University. Both experimenters had experience conducting experiments, and both had some experience with computers. The experimenters were blind to the hypotheses guiding the experiments. The experimenters were trained in using the testbed and conducting the experiment by two DSC team members: a domain expert (Major, U.S. Army Reserve) and a Ph.D. psychologist. The experimenters were trained the day before the experiment began; DSC personnel remained for 1 1/2 additional days (i.e., after the experiment began) to help ensure that the experiment was being conducted as designed.

The session with each participant began with them reading the "your job" description in Appendix A which described the basic air defense task as represented in the testbed and described in Section 3.2 of this report. After they had completed reading the description, the experimenter discussed it with them to make sure they understood it, particularly the differences between the testbed and an actual air defense system.

The order in which the participants were tested on the different conditions of the experiment was counterbalanced as shown in Table 3-7. As can be seen, participants worked the high- and low-workload conditions for a level of the Person-Machine Interface factor before moving on to another interface level. The project team made this decision as a result of pilot testing in order to facilitate the training of each interface condition. After all participants said that they fully understood the "your job" description, they were given a written description of their first interface condition. After the participant finished reading the description, the experimenter working with the participant would call-up a training run of the interface condition and demonstrate the key points of the description. Then, the participants had an opportunity to work with the interface condition until they said they felt comfortable with it. The first familiarization session took approximately 30 minutes because the participants had to take time familiarizing themselves with the characteristics of the testbed, particularly using the mouse for all interaction with the system. Subsequent familiarization sessions for the other two interface conditions took approximately 15 minutes.

After each familiarization session, participants had a practice session simulating an actual test session. Targets appeared on the screen every 8 seconds, thereby representing a workload level halfway between the actual high-and low-workload conditions. Participants were urged to per-

form as well as they could, but not to hesitate in asking the experimenter questions about the system's capabilities. The experimenters observed participants' performance during the practice session, pointing out only aspects of the written system descriptions that the participants seemed not to be considering when performing the air defense task. The practice session for all three interface conditions took approximately one-half hour.

Table 3-7

Experiment 1: The Order of Experimental Conditions Used for Counterbalancing

Participant*	Session (in Time Order) 1	2	3	4	5	6
13	MH1	ML2	OL3	OH2	SH3	SL1
14	OH1	OL2	SL3	SH2	MH3	ML1
9**	SH1	SL2	ML3	MH2	OH3	OL1
4	ML2	MH3	OH1	OL3	SL1	SH2
5	OL2	OH3	SH1	SL3	ML1	MH2
12	SL2	SH3	MH1	ML3	OL1	OH2
10	SH3	SL1	OL2	OH1	MH2	ML3
11	OH3	OL1	ML2	MH1	SH2	SL3
6	MH3	ML1	SL2	SH1	OH2	OL3
15	SL3	SH2	OH1	OL2	ML1	MH3
7	OL3	OH2	MH1	ML2	SL1	SH3
8	ML3	MH2	SH1	SL2	OL1	OH3
1	MH2	ML1	OL3	OH1	SH3	SL2
2	OL2	OH1	SH3	SL1	ML3	MH2
3	SH2	SL1	ML3	MH1	OH3	OL2

M - Manual
O - Override
S - Screening

L - Low Workload
H - High Workload
1,2,3 - Target Orderings

*Participants were matched randomly with rows of the table.
**Due to an error, data for Participant 9 were lost.

After answering any questions, the experimenter would then start the designated test condition. ("Training to criterion" was preferred but abandoned prior to the experiment because the project team was uncertain as to whether it could be implemented in a manner ensuring that every participant completed all six cells of the design within the eight-hour time limit.) After completing the test condition, the participant filled out a questionnaire asking the participant to evaluate his/her performance, the workload level, and the degree to which s/he liked working with the system. After completing the questionnaire, the experimenter started the second test case with the interface condition, but now with the other workload level. The sequence of "read system description, familiarization session, practice session, first interface test condition, questionnaire, second test condition, and questionnaire" was used for all three interface

conditions. Appendix B contains the "system description" for all th... interface conditions.

3.4.1.6 *Dependent Measures*. There were three sets of dependent measures in the experiment: (1) objective performance measures; (2) objective workload measures; and (3) subjective performance, workload, and preference measures. The measurement of participants' performance was straightforward because the true identity of each target was known. Moreover, given the manner in which the 200-target simulation was constructed, it was possible to examine performance for different groups of targets. Specifically, performance for the different experimental conditions was examined for: (a) the entire 200-target simulation; (b) the 85 targets with a degree of belief $\geq .80$; (c) the 57 targets with a degree of belief between .60 and .80; and (d) the 58 targets with a degree of belief below .60 before participants collected additional information via IFF, HQ ID, or messages. Groups (b), (c), and (d) represent, respectively, the "firmly identified," "questionably identified," and (d) "unknowns" classifications used in the "screening" interface. Group (d) is most comparable to the "p=.5 without the extra cue" targets analyzed in detail in Phase I.

Two objective workload measures depended on the secondary task performed by the participants. The first objective workload measure was the participants' response time to the "response light." The second measure was their response accuracy; participants were told to press the middle button of the mouse when the light was red and the right-hand button of the mouse when the light was green.

Two questionnaires were used to obtain participants' subjective performance, workload, and preference measures. One questionnaire was given after the completion of each of the six test cases (i.e., one for each condition in the design) and asked participants to rate (on a 9-point scale) their performance, the level of workload, and the degree to which they liked working with the system. The second questionnaire was given at the end of the experimental session, immediately after the participants completed the questionnaire after the sixth test case. Again, participants were asked to rate on a 9-point scale how well they think they performed with each of the three systems, how hard they worked to perform the aircraft identification task with each system, and how much they liked working with each system. In addition, each participant rated how much each cue affected his/her ability to get correct identifications so that we would have a subjective measure of what information the participants considered useful. The cues were HQ ID, IFF, pop-up, corridor, speed, altitude, distance, heading, jammer, and messages. Copies of both questionnaires can be found in Appendix C.

3.4.2 Results. This section of the report presents the results, in turn, for each of the three sets of dependent measures.

3.4.2.1 *Performance*. The performance results are organized according to the target's degree of belief (in Shaferian terms) before participants collected additional information via IFF, HQ ID, or messages. In particular, Table 3-8 presents the "mean percent correct" for each human-

machine interface factor for each of the two workload levels, for the N=58 targets that had a degree of belief (in Shaferian terms) below .60 (on a 1.0 scale) before participants collected additional information.

Table 3-8

Experiment 1: The "Mean Percent Correct" for N=58 Targets with Initial Degree of Belief Below .60

		Workload Factor		
		Low	High	$\bar{X}$
Human-Machine Interface Factor	Manual	69.8	24.5	47.2
	Override	72.2	53.2	62.7
	Screening	75.5	47.0	61.3
	$\bar{X}$ =	72.5	41.6	57.1

These data are most comparable to the "p=.5 without the extra cue" data analyzed in detail in Phase I. The principal points to make on the basis of repeated measures, ANOVA and paired comparisons, are as follows:

1. In contrast with Phase I we obtained a significant Main Effect for the human-machine interface factor due to significantly better performance in override and screening than in the manual condition [$\bar{X}_M$=47.2, $\bar{X}_O$=62.7, $\bar{X}_S$=61.3, SS_H=.402, SS_E=.351, F(2,26)=14.894, p<.001].

2. As in Phase I, we obtained a significant Main Effect for workload [$\bar{X}_L$=72.5, $\bar{X}_H$=41.6, SS_H=5.972, SS_E=.385, F(1,13)=201.576, p<.001]. In contrast to Phase I, this difference was significant for each interface condition.

3. And, as in Phase I, we obtained a significant interaction. This was primarily due to extremely low performance in the manual-high workload condition ($\bar{X}$=24.5). As in Phase I, we found no significant differences between the three conditions under low workload. In contrast to Phase I, however, we found no significant difference in performance in the override and screening conditions under high workload. Both conditions were significantly better than the manual-high workload condition [$\bar{X}_{M,H}$=24.5, $\bar{X}_{O,H}$=53.2, SS_H=1.123, SS_E=.058, F(1,13)=251.239, p<.001; $\bar{X}_{M,H}$=24.5, $\bar{X}_{S,H}$=47.0, SS_H=.688, SS_E=.235, F(1,13)=38.094, p<.001]. And, in contrast to Phase I, both conditions under high workload resulted in significantly worse performance than that achieved in the manual-low workload condition [$\bar{X}_{M,L}$=69.8, $\bar{X}_{O,H}$=53.2, SS_H=.387, SS_E=.521, F(1,13)=9.651, p=.008; $\bar{X}_{M,L}$=69.8, $\bar{X}_{S,H}$=47, SS_H=.727, SS_E=.645, F(1,13)=14.639, p=.002].

Tables 3-9 and 3-10 present the "mean percent correct" for the (a) N=57 targets with an initial degree of belief between .6 and .8, and (b) the total N=115 targets with an initial degree of belief below .8, respec-

tively. All the results presented above for the ANOVAs and paired comparisons replicated those for the N=58 targets with an initial degree of belief $\leq .60$ with two exceptions. Specifically, the manual-low workload condition was (a) *not* significantly higher than that for the override or screening interfaces under high workload for the N=57 targets with degrees of belief between .6 and .8, and (b) *not* significantly higher than override-high workload condition, but significantly higher than the screening-high workload condition [$\bar{X}_{M,H}=70.8$, $\bar{X}_{S,H}=57.2$, $SS_H=.259$, $SS_E=.43$, $F(1,13)=7.833$, $p=.015$] for the N=115 targets with an initial degree of belief < .80.

Table 3-9

Experiment 1: The "Mean Percent Correct" for N=57 Targets with Initial Degree of Belief Between .6 and .8

		Workload Factor		
		Low	High	$\bar{X}$
Human-Machine	Manual	71.8	36.7	54.25
Interface	Override	77.7	69.3	73.50
Factor	Screening	77.7	67.5	72.6
	$\bar{X}$ =	75.7	57.8	66.78

Table 3-10

Experiment 1: The "Mean Percent Correct" for N=115 Targets with Initial Degree of Belief Below .80

		Workload Factor		
		Low	High	$\bar{X}$
Human-Machine	Manual	70.8	30.7	50.75
Interface	Override	74.9	61.2	68.05
Factor	Screening	76.6	57.2	66.90
	$\bar{X}$ =	74.1	49.7	61.90

As indicated in Section 3.4.1.4, 28 of the 58 "unknown" targets and 28 of the 57 "questionable" targets could be perfectly identified without IFF challenges or requests for HQ ID. This was accomplished by having the messages focus on groups of targets that were in the safe-passage corridor, but that had particular values for pop-up, speed, and altitude. This action was taken in an attempt to make participants have a reason for using this information and not simply perform the identification task solely on the basis of visual information (i.e., being out of the corridor and/or jamming) or HQ ID and IFF challenges.

Table 3-11 presents the "mean percent correct" for the N=56 targets that could be perfectly identified from messages without IFF challenges or requests for HQ ID. Statistical tests for this subset of targets replicated the results presented above for the N=58 targets with an initial degree of belief $\leq .6$.

Table 3-11

Experiment 1: The "Mean Percent Correct" for "Message Targets"

		Workload Factor		
		Low	High	$\bar{X}$
Human-Machine	Manual	70.1	33.4	51.75
Interface	Override	73.4	58.9	66.15
Factor	Screening	75.5	55.7	65.60
	$\bar{X}$ =	73.0	49.33	61.17

Tables 3-12 and 3-13 present the "mean percent correct" for (a) the N=85 targets with an initial degree of belief ≥ .80, and the (b) entire N=200 target simulation, respectively.

Table 3-12

Experiment 1: The "Mean Percent Correct" for N=85 Targets with Initial Degree of Belief ≥ .80

		Workload Factor		
		Low	High	$\bar{X}$
Human-Machine	Manual	85.0	36.1	60.55
Interface	Override	93.2	88.8	91.00
Factor	Screening	90.2	87.1	88.65
	$\bar{X}$ =	89.5	70.7	80.07

Table 3-13

Experiment 1: The "Mean Percent Correct" for the Entire 200-Target Simulation

		Workload Factor		
		Low	High	$\bar{X}$
Human-Machine	Manual	76.8	33.0	54.90
Interface	Override	82.7	72.9	77.80
Factor	Screening	82.4	69.9	76.15
	$\bar{X}$ =	80.63	58.6	69.62

Again, we obtained a significant main effect for workload, with low workload resulting in better performance than high workload for each of the three human-machine interface conditions. Again, we obtained a significant main effect for human-machine interface conditions, with override and screening both resulting in significantly higher performance than the manual condition for both analyses. And, we also obtained the previous significant workload × interface interactions. Again, we found that both the override and screening conditions resulted in significantly higher performance than the manual condition in only the high-workload condition. In addition, however, in contrast to the previous analyses, we also found that

the override condition resulted in significantly higher performance than the screening condition under high workload for both the N=85 target set [$\bar{X}_{O,H}$=88.8, $\bar{X}_{S,H}$=87.1, SS_H=.004 SS_E=.007, F(1,13)=7.937, p=.015] and for the entire N=200 target simulation [$\bar{X}_{O,H}$=72.9, $\bar{X}_{S,H}$=69.9, SS_H=.013 SS_E=.025, F(1,13)=6.61, p=.023]. In both cases, however, the performance achieved in the override and screening conditions under high workload were not significantly different than that achieved in the manual condition under low workload.

Table 3-14 presents the percentage of targets correctly identified by a totally automated system using only the initially available information for different classes of targets. Examination of Tables 3-8 through 3-14 and repeated-measures, multivariate tests using the totally automated system's "percent correct" score as the null hypothesis showed two principal findings. First, using the manual interface, participants were able to perform as well and, more often than not, better than a totally automated system under low workload. However, they performed significantly worse than a totally automated system under high workload for all comparisons. Second, using either the override or screening interface, participants performed significantly better than the totally automated system for all comparisons under low workload. They performed as well as the totally automated system under high workload. The override interface also resulted in significantly better performance than the totally automated system using only the initially available information for all comparisons under high workload. However, under high workload, the screening condition significantly outperformed only the totally automated system for targets with an initial degree of belief between .6 and .8. The latter results are in sharp contrast to those obtained in the Phase I experiment. They suggest that the participants were not able to effectively utilize additional information in the high-workload condition when using the initial implementation of the screening interface for a highly representative air defense scenario.

Table 3-14

Percent Correct for a Totally Automated System (i.e., with Initial Data Only) for Different Classes of Targets

Initial Degree of Belief $\leq$ .6	= 46.55
.6 < Initial Degree of Belief < .8	= 64.28
Initial Degree of Belief $\geq$ .8	= 87.33
Entire 200-Target Simulation	= 69.00

3.4.2.2 *Workload.* Two objective workload measures were developed tied to the secondary task performed by the participants. The results for each measure are presented, in turn.

The first objective measure was the participants' response time to the response light. Although ANOVAs found no significant main effects or interactions, planned comparison tests did find that participants took significantly longer, on the average, to respond to the light using the over-

ride interface when in the low- than high-workload condition [$\bar{X}_{O,L}$=1256.7, $\bar{X}_{O,H}$=1187.7, SS_H=66654, SS_E=65426, $F(1,13)$=13.24, p=.003].

The second objective workload measure was the accuracy of the participants' mean responses. ANOVAs and paired comparisons found no significant differences in the mean response accuracies for the different conditions.

3.4.2.3 *Subjective Measures.* Two questionnaires were used to obtain participants' subjective performance, workload, and preference measures. The results for each questionnaire are considered, in turn.

The first questionnaire was given immediately after the completion of each condition in the design. Participants were asked to rate, on a 9-point scale, their performance, the level of workload, and the degree to which they liked working with the system. Table 3-15 presents the mean responses for subjective performance, workload, and preference for each of the six conditions in the design. Higher values represent better scores on all three subjective measures.

Table 3-15

Experiment 1: Mean Subjective Performance, Workload, and Preference (First Questionnaire)

(a) *Subjective Performance*

		Workload Factor		
		Low	High	$\bar{X}$
Human-Machine Interface Factor	Manual	5.67	3.73	4.70
	Override	6.53	6.47	6.50
	Screening	6.60	6.33	6.47
	$\bar{X}$ =	6.27	5.51	5.89

(b) *Subjective Workload*

		Workload Factor		
		Low	High	$\bar{X}$
Human-Machine Interface Factor	Manual	5.33	3.47	4.40
	Override	5.40	4.73	5.07
	Screening	5.93	4.93	5.43
	$\bar{X}$ =	5.55	4.38	4.97

(c) *Subjective Preference*

		Workload Factor		
		Low	High	$\bar{X}$
Human-Machine Interface Factor	Manual	5.87	4.67	5.27
	Override	5.93	5.60	5.77
	Screening	6.00	5.73	5.87
	$\bar{X}$ =	5.93	5.33	5.63

A repeated-measures ANOVA on the subjective performance ratings replicated the significant main effects and interface x workload interac-

tion obtained for the objective performance. Specifically, on the average, participants thought they did worse in the high- than low-workload condition [$\bar{X}_L$=6.27, $\bar{X}_H$=5.51, SS_H=38.53, SS_E=.115.47, F(1,14)=4.672, p=.048], and with the manual than override or screening interfaces [$\bar{X}_M$=4.7, $\bar{X}_O$=6.5, $\bar{X}_S$=6.47, SS_H=63.62, SS_E=108.71, F(2,28)=8.193, p=.002]. In particular, participants thought they performed extremely poorly in the manual-high workload condition, as, in fact, was the case. However, participants also thought they performed significantly worse with the manual interface under the low-workload condition than with the override and screening interfaces under either low or high workload; this was, in fact, not the case.

An ANOVA on the subjective workload ratings shows that participants thought it was easier to perform the task under the low-workload condition [$\bar{X}_L$=5.55, $\bar{X}_H$=4.38, SS_H=93.63, SS_E=117.87, F(1,14)=11.12, p=.005]. The main effect for interface was not significant; however, the interface × workload interaction was significant [SS_H=5.76, SS_E=22.24, F(2,28)=3.622, p=.04], because of the great difficulty the participants perceived in the manual-low workload condition.

The subjective workload results contrast sharply with the objective workload results. This could be due to expected differences between objective and subjective workload measures. However, it might also be the result of our secondary task being perceived as part of the primary task by our participants because they gained or lost points on the basis of their performance in acknowledging orders from higher headquarters.

An ANOVA also was performed on the participants' preference ratings. The only significant effect was a main effect for workload; participants preferred working in the low- than high-workload condition [$\bar{X}_L$=5.93, $\bar{X}_H$=5.63, SS_H=30.0, SS_E=75.0, F(1,14)=5.6, p=.033]. Paired comparisons showed that participants preferred working in each of the five other conditions significantly more than they preferred working in the manual-low workload condition.

The second questionnaire was given at the end of the experimental session, immediately after the participants completed the questionnaire after the sixth test case. Again, participants were asked to rate on a 9-point scale how well they think they performed with each of the three systems, how hard they worked to perform the aircraft-identification task with each system, and how much they liked working with each system. ANOVAs found no significant differences in the mean values for the three systems for the subjective performance workload, or preference ratings. This result, when taken in conjunction with the results for the first questionnaire, demonstrates the importance of the workload factor in participants' subjective opinion of the systems.

The second questionnaire also asked participants to rate how much each cue affected their ability to obtain correct identification, thus providing a subjective measure of cue utility. The mean cue utility measures are shown in Table 3-16.

Table 3-16

Mean Utility Rating for the Target Identification Cues

Cue		Mean	Cluster
IFF	=	8.14	Cluster 1
HQ ID	=	7.39	
Corridor	=	6.75	Cluster 2
Speed	=	5.89	
Altitude	=	5.75	
Messages	=	5.57	
Jamming	=	5.07	
Pop-Up	=	4.36	Cluster 3
Distance	=	3.35	
Heading	=	3.35	

Paired comparison tests suggest that the cues can be grouped into roughly three clusters. Specifically, the first cluster would include IFF and HQ ID. The mean values for these two cues were not significantly different, yet they were both significantly higher than all other cues at the $p<.05$ level. The second cluster would include corridor, speed, altitude, messages, and jamming. The mean values for these cues were not significantly different, although the comparisons for corridor versus speed, altitude, and messages approached significance (i.e., $<.10$). The third cluster comprises pop-up, distance and heading. In almost all cases, the mean values for the cues in the second cluster were significantly higher than the mean values for the cues in the third cluster.

3.4.3 Discussion. In total, the conditions in the first experiment supported the hypotheses in most, but not all, cases. Specifically, as in Phase I, we obtained a significant main effect for workload. For all classes of targets, participants' mean performance was better in the low- than high-workload condition. And for all classes of targets, we obtained a significant main effect for the human-machine interface factor due to significantly better mean performance with override and screening than manual. And, as in Phase I, we obtained a significant interaction due primarily to the extremely low performance in the manual-high workload condition. However, in sharp contrast to Phase I when screening was used in the high-workload condition, participants were unable to maintain the level of performance they achieved under the low-workload condition. Participants had higher mean performance scores in the override-high workload condition than in the screening-high workload condition for all groupings of targets on the basis of the initial degree-of-belief values, as well as for the entire 200-target simulation. For the N=65 targets with degree of belief > .80, the mean values for override ($\bar{X}_O$=88.8) and screening ($\bar{X}_S$=87.1) under high workload were extremely similar, but significantly different because of the extremely small error variance (sums of squares for error term equaled .007). In contrast, we think the significantly higher mean value for override (72.9) versus screening (69.9) in the high-workload condition for the 200-target test case was more a function of (large) sample size than (small) error variance.

It is important to note here that the two DSC team members who trained the experimenters at the beginning of the first experiment raised the possibility with the project team that there might be no difference between override and screening in high workload because they observed that a small percentage of the targets were leaving the screen as "unknown" in the screening condition. We think this occurred because the task in the Fort Bliss experiment was more difficult than the Phase I task, and in the screening condition, there was a higher proportion of "unknowns" in the Fort Bliss than Phase I simulation. Consequently, the screening condition in the second experiment was modified so that no targets left the screen as "unknown;" if the participant did not have time to identify the target, the computer classified it as Friend or Hostile, depending on which hypothesis had the highest degree of belief, just as in the override condition. This is obviously a reasonable thing to do for generalizing to a real combat system. More importantly, it permits a better test of the value of override versus screening in facilitating participants' information processing and, in turn, performance under high-workload conditions.

We now turn to discuss the second Fort Bliss experiment. In addition to the modified screening interface, the second experiment tested the utility of an allocation (i.e., rule-creation) capability under the high-workload condition.

3.5 Experiment 2

This section of the report describes the method and results of the second experiment conducted at Fort Bliss, Texas.

3.5.1 Method. The method section for Experiment 2 is composed of the following subsections: the experimental design, the participants, how the factors in the design were operationalized, a description of the air defense simulation, the procedures used when conducting the experiment, and the dependent measures used to assess the effects of the different interface conditions. Each subsection is considered, in turn.

3.5.1.1 *Experimental Design*. The second experiment was a within-subject, repeated-measures design with the following five conditions, all under high-workload only: completely manual with allocation, override without allocation, override with allocation, screening without allocation, and screening with allocation. A completely automated condition where the system performed the target identification task without access to IFF, HQ ID, or message data served as the baseline condition.

3.5.1.2 *Participants*. Fourteen U.S. Army air defense operators participated in the experiment between 2 June - 19 June 1987. All participants were either first or second lieutenants who had completed the Basic Course and who had some experience with either the PATRIOT or HAWK air defense system. None of the participants had participated in Experiment 1.

3.5.1.3 *Operationalizing Design Factors*. High workload was again operationalized by having a target appear on the screen every 4 seconds, as

in Experiment 1. This resulted in a total of 355 targets being displayed on the screen during the course of the session. As in Experiment 1, the 200 performance targets were sandwiched between 25 leading targets to get participants fully engaged in the task and 130 trailing targets that ensured that all performance targets had left the display before the simulation ended. The same three differently ordered files of the targets that were used in Experiment 1 to minimize the effects of memory on performance were also used in Experiment 2.

Except for one relatively minor change, the override:no allocation condition was the same as that used in Experiment 1. The system made an initial identification for all incoming targets (black circles for friends and black diamonds for foes) on the basis of (a) whether and where it popped up; (b) whether the target was in the corridor or not; (c) whether its speed and altitude met the corridor parameters if it was in the corridor; and (d) whether or not the target was a jammer. Participants could override (or change) the target identification on the basis of available and newly collected information, that is, IFF responses, HQ identifications, and messages. Again, responses to an IFF challenge went directly into the system to maintain representativeness with the actual air defense domain; consequently, it also was possible for the system to change its identification on the basis on this new information. The one change was that the system no longer changed an aircraft identification made by the participant if the aircraft subsequently moved into or out of the corridor. While this "change in identification" happened seldom in Experiment 1 (e.g., ≤ 5 times per session, on the average), participants found it particularly disconcerting when the system changed an identification that the subject made on the basis of "HQ ID" information, which the system did not have.

The screening:no allocation condition was basically the same as that used in Experiment 1. When, on the basis of the initially available information (which was the same information initially available in the override condition), the target had a degree of belief $\geq .80$, the target identified the target as "firmly identified; foes were red diamonds and friends were blue circles. When the target had an initial degree of belief between .60 and .80, the system classified the target as either a "questionable friend" (circle) or a "questionable foe" (diamond), and indicated whether this was due to uncertainty, by using the color black, or conflict, by using the color purple. If the system was unable to identify the target because of either uncertainty or conflict (i.e., degree of belief <.60), the system classified the target as unknown (a black circle). The "highest priority unknown" was identified by a purple, solid circle around it.

There were, however, two changes in the screening condition in Experiment 2. First, the screening condition in the second experiment was modified so that no targets left the screen as "unknown;" if the P did not have time to identify the target, the computer classified it as friend or foe depending on which hypothesis had the highest degree of belief, just as in the override condition. And, second, as in the override condition, the system no longer changed an identification made by the participant.

In the manual:allocation condition, the participants were still responsible for identifying all targets, but now they could "allocate" some of the decision making to the system by creating rules that the system could use to identify incoming aircraft. For example, participants (Ps) could tell the system to identify all jammers as hostile, and the system would do so automatically. Or, for example, Ps could tell the system to automatically identify all aircraft in a safe-passage corridor with the correct speed and altitude as friends, and so forth.

The allocation capability worked the same in the manual:allocation, override:allocation, and screening:allocation conditions. Before beginning an attack phase, the Ps had the opportunity to create identification rules. The system started off with the display shown in Figure 3-4. Table 3-17 identifies the options for each button in the rule-creation component of the system. For example, the POPUP button had four options: No, Close, Feba, and Yes. The CORR button had five options: In, One-Out, Two-Out, Out, and N/A. Let us assume, for example, that the P wanted to say that all jammers were to be identified as hostile. The P would move the mouse to the JAMM button and click the left mouse button. When YES came up, indicating that the P was referring to jammers, the P would then go over to the RESULT column in the far left-hand corner of the display and click-on "Hostile," implying that the P wanted all jammers to be identified as hostile. Then, the P would click-on "Save Rule," which is directly below the "Hostile" button, to save the identification rule. The system now understands that all jammers are to be identified automatically as hostile.

If Ps now clicked on RULES at the top of the display, making sure to hold down the mouse button, they would see that a rule called R6(H) had been created. At any time, the Ps could move the mouse over R6(H) and lift their finger off the mouse button. The values of the R6(H) identification rule would then appear on the screen; that is, "jammers are hostile." If Ps wanted to erase this rule, they need to click-on "Clear Rule" directly below "Save Rule." As another example, if Ps wanted all aircraft doing everything correct in the corridor to be identified as friend, they would do the following:

- click-on CORR until it read In;
- click-on Friend; and
- click-on Save Rule.

Before creating a new rule, Ps needed to select a rule number. To do this, they (1) clicked-on RULES, and (2) while pressing the mouse button, moved the mouse to an empty rule and released the mouse button. Initially, rule R6 was selected for the P because the first five spaces for saving rules in the RULES box were listed under MESSAGE. Message rules were executed first by the system because they indicated rules that were created to identify specific groups of aircraft with the characteristics indicated in messages. For example, assume Ps received the message below:

(1) Hostile Group Profile
(1) Popup=No: Corr=TwoOut
(1) Alt=11,000: Speed=1200

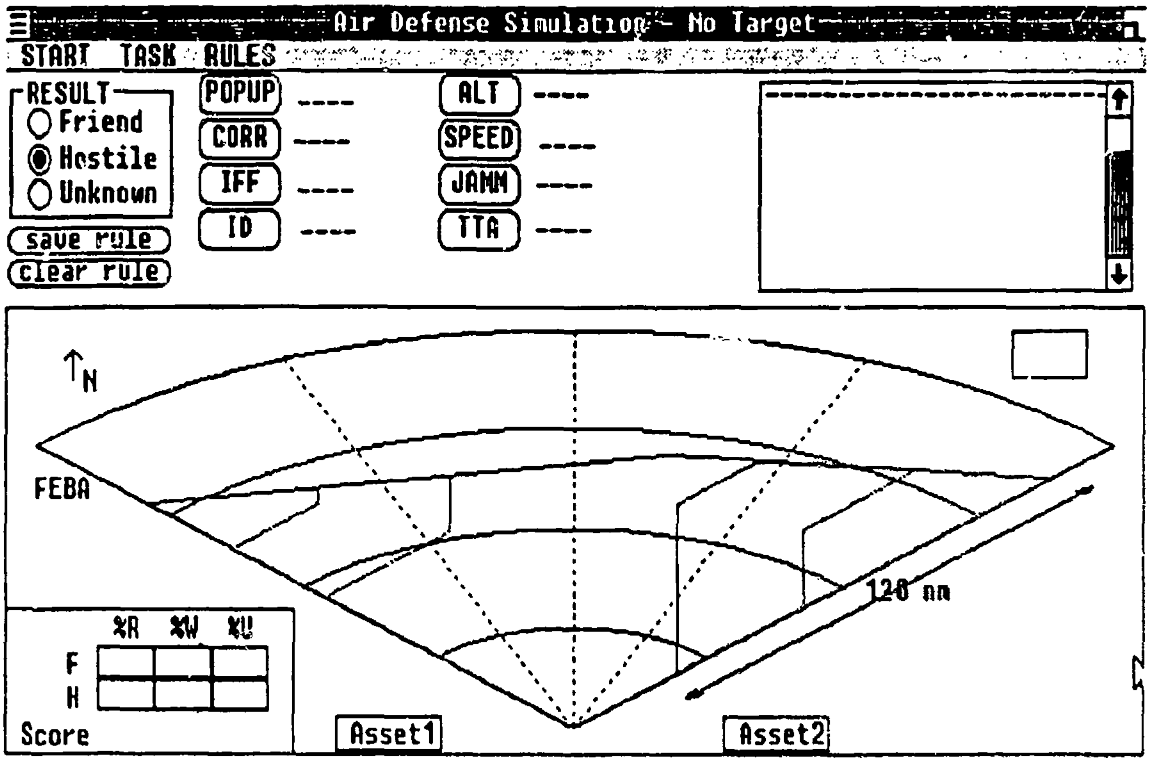


Figure 3-4. Initial screen display.

Table 3-17 M,O

Rule Work Sheet

Rule # ___

ID as:	FRIEND HOSTILE UNKNOWN	

<u>Criteria</u>

POPUP :	NO	(Target did not pop up)
	CLOSE	(Target popped up close to friendly asset)
	FEBA	(Target popped up near FEBA)
	YES	(Target popped up close *or* near FEBA)
CORR :	IN	(Target in corridor, speed *and* altitude are correct)
	ONE OUT	(Target in corridor, speed *or* altitude incorrect)
	TWO OUT	(Target in corridor, speed *and* altitude incorrect)
	OUT	(Target not in corridor)
	N/A	(Targets have not reached corridor entrance)
IFF :	POSITIVE	(Targets that *have* been challenged, respond as FRIEND)
	NEGATIVE	(Targets that *have* been challenged, no response)
	NO CHAL	(Targets that have not been challenged)
ID:	UNKNOWN	(All unknown targets)
ALT: (altitude band, feet)		0:1000 1000:2000 2000:5000 5000:10000 10000:80000 80000+
SPEED : (knots)		0:200 200:400 400:600 600:800 800:1000 1000+
JAMM :	YES	(All targets that are jamming)
	NO	(All targets that are not jamming)
TTA : (Time to Asset)	1 MIN	(All targets within 1 minute from friendly assets)
	2 MIN	(All targets within 2 minutes from friendly assets)
	3 MIN	(All targets within 3 minutes from friendly assets)
	3 MIN+	(All targets more than 3 minutes from friendly assets)

They could create a rule that identified all these aircraft as hostile by doing the following:

- select Rule MR1;
- click-on POPUP until it read No;
- click-on CORR until it read TwoOut;
- click-on ALT until it read 10000-80000;
- click-on SPEED until it read 1000+;
- click-on Hostile; and
- click-on Save Rule.

This rule would be stored as MR1(H) in the Message category under the RULES menu. It would identify all aircraft as hostile with the above characteristics. Ps needed to erase or 'clear' this rule from the RULES menu when this message (i.e., Message 1) disappeared from the message box because all the hostile aircraft with these (message) characteristics had left the screen.

Ps could create as many rules as they liked before and during an attack phase. The message rules always were executed first. The other rules were executed in the order that they were listed. It was emphasized that this order was important. For example, if Ps wanted all jammers to be hostile and any other aircraft that were traveling correctly within a corridor to be friendly, then the rule for jammers had to come before the corridor rule. (Note: Although this sounds complicated, it was readily communicated during training.)

In the manual:allocation condition, any target not covered by a rule was identified as unknown; a target not covered by a rule in the override:allocation condition was identified as friend or foe on the basis of the initial information; and, in the screening condition, identification was based on the degree of belief (i.e., $\leq .6$, between .6 and .8, and $\geq .8$) for the initial information. (The written description of these three conditions given to the participants is found in Appendix D of this report.)

3.5.1.4 *Simulation*. The air defense simulation used in the high-workload condition in Experiment 1 was used in Experiment 2. The simulation is discussed in detail in Section 3.4.1.4 of this report.

3.5.1.5 *Procedures*. The same basic procedures as those used in Experiment 1 were used in Experiment 2. Each of the fourteen participants took an entire workday (approximately 8 hours, not including a one-hour lunch break) to perform the air defense task for each of the five interface conditions in the design. Most of the time, two participants participated each day the experiment was conducted at the ARI Field Unit's offices at Fort Bliss. The participants were separated by partitions, and they worked independently on identical testbed systems. In addition, there were two experimenters, one for each participant.

The session with the participants began with them reading the "your job" description in Appendix A which described the basic air defense task as represented in the testbed and described in Section 3.2 of this report. After they had completed reading the description, the experimenter dis-

cussed it with them to make sure they understood it, particularly the differences between the testbed and an actual air defense system.

The order with which the participants were run through the different conditions of the experiment was counterbalanced as shown in Table 3-18. After all participants said that they fully understood the "your job" description, they were given a written description of their first interface condition. After the participant finished reading the description, the experimenter working with the participant would call-up a training rule of the interface condition and demonstrate the key points of the description. Then, the participants had an opportunity to work with the interface condition until they said they felt comfortable with it. The first familiarization session took approximately 30 minutes because the participants had to take time familiarizing themselves with the characteristics of the testbed, particularly using the mouse for all interaction with the system. Also, the first session with an "allocation condition" took approximately 30 minutes of training. Subsequent familiarization sessions for the other interface conditions took approximately 15 minutes.

Table 3-18

Experiment 2: The Order of Experimental Conditions Used for Counterbalancing

	Session (in Time Order)				
Participant*	1	2	3	4	5
5	MA2	SA3	SN1	ON3	OA2
11	ON2	OA3	MA1	SA3	SN2
3	SN2	SA3	OA1	ON3	MA2
14	MA1	SN2	SA3	OA2	ON1
9	OA1	ON2	MA3	SN2	SA1
15	SN1	SA2	OA3	ON2	MA1
2	MA3	SA1	SN2	ON1	OA3
1	ON3	OA1	MA2	SA1	SN3
7	SA3	SN1	ON2	OA1	MA3
13	SA2	SN3	MA1	ON3	OA2
10	ON2	OA1	SA1	SN3	MA2
12	MA2	ON3	OA1	SA3	SN2
4	SN1	SA2	MA3	OA2	ON1
6	OA1	ON2	SN3	SA2	MA1
8**	MA1	OA2	ON3	SN2	SA1

M - Manual
O - Override
S - Screening

A - Allocation
N - No Allocation
1,2,3 - Target Orderings

*Participants were matched randomly with rows of the table.
**Only 14 of the 15 requested participants were available for the experiment. The sessions were not conducted for "Participant 8" to minimize counterbalancing effects.

After each familiarization session, participants had a practice session simulating an actual test session. Targets appeared on the screen every 4 seconds, just like the test sessions. Participants were urged to perform as well as they could, but not to hesitate in asking the experimenter questions about the system's capabilities. The experimenters observed participants' performance during the practice session, pointing-out only aspects of the written system descriptions that the participants seemed not to be considering when performing the air defense task. The practice session for each interface condition took approximately one-half hour.

After answering any questions, the experimenter would then start the designated test condition. After completing the test condition, the participant filled out a questionnaire asking the participant to evaluate his/her performance, the workload level, and the degree to which s/he liked working with the system. After completing the questionnaire, the experimenter would start the second interface condition. The sequence of "read system description, familiarization session, practice session, first interface test condition, questionnaire, second test condition, and questionnaire" was used for all three interface conditions.

The same two experimenters who participated in Experiment 1 also participated in Experiment 2. The experimenters were trained in how to operate the three allocation conditions by a DSC team member the day before the experiment began. The DSC team member remained for an additional day (i.e., after the experiment began) to help ensure that the experiment was being conducted as designed. As in Experiment 1, the experimenters were blind to the hypotheses guiding the experiment.

3.5.1.6 *Dependent Measures*. There were four sets of dependent measures in the experiment: (1) objective performance measures; (2) objective workload measures; (3) subjective performance, workload, and preference measures; and (4) measures of participants' information-processing strategies. The first three sets of dependent measures were the same as those used in Experiment 1, and are not discussed here. The information-processing measures included (a) the percentage of targets hooked; (b) the length of hooking (in seconds); (c) the order in which cues were used; and (d) the types of rules participants created in the allocation conditions.

It is important to note that the information-processing analyses were a long, slow time-consuming process. *All* of the participants' mouse clicks while performing the experiment were stored in data files on the IBM-AT testbeds. These data files had to be reduced for analysis. This required the development of some computer programs for data reduction, the subsequent transfer of the reduced data files to spreadsheets for formatting and, finally, their input into the statistical package. Therefore, due to time and resource constraints, the information-processing analyses were performed only for Experiment 2. This experiment was selected because it included the allocation conditions in addition to the basic, override and screening condition. Consequently, the second Fort Bliss experiment represented the best situation for proposing information-processing

guidelines for future applications based on our theoretical position and empirical results.

3.5.2 Results. This section of the report presents the results, in turn, for each of the four sets of dependent measures.

3.5.2.1 *Performance*. The performance results are organized according to the target's degree of belief (in Shaferian terms) before participants collected additional information via IFF, HQ ID, or messages. In particular, Table 3-19 presents the "mean percent correct" for each human-machine interface condition for the N=58 targets that had a degree of belief (in Shaferian terms) below .60 (on a 1.0 scale) before participants collected additional information. In addition, Table 3-19 presents the mean performance score for the high workload-manual condition from the first Fort Bliss experiment for comparison purposes.

Table 3-19

Experiment 2: The "Mean Percent Correct" for N=58 Targets with Initial Degree of Belief ≤ .60

		Allocation (i.e., Rule-Creation) Capability	
		No Allocation Capability	Allocation Capability
Human-Machine Interface Factor	Completely Manual	24.5*	57.4
	Override	59.6	69.7
	Screening	74.2	75.6

*Data from High-Workload Manual Condition in Experiment 1

The principal results are presented below. Paired comparison tests were performed only for the five conditions in Experiment 2.

1. Examination of Table 3-19 shows that all five conditions in Experiment 2 markedly improved performance over that achieved in the manual: high-workload condition in Experiment 1. Moreover, all five conditions performed better than the totally automated system (percent correct = 46.55) for the set of targets with an initial degree of belief ≤ .60.

2. In contrast with Experiment 1, screening without allocation resulted in significantly higher performance for targets with initial degrees of belief ≤ .60 than did override without allocation [$\bar{X}_{O:NA}=59.6$, $\bar{X}_{S:NA}=74.2$, $SS_H=.306$, $SS_E=.098$, $F(1,13)=40.557$, $p<.001$]. This indicates that the statistically equivalent level of performance obtained in Experiment 1 was a function of the screening interface permitting targets to leave the display classified as "unknown" instead of classifying the targets on the basis of available information, as in Experiment 2. Moreover, screening without allocation resulted in significantly better performance than that achieved in the manual

with allocation condition [$\bar{X}_{S:NA}$=74.2, $\bar{X}_{M:A}$=57.4, SS_H=.404, SS_E=.481, $F(1,13)$=10.930, p=.006].

3. Having an allocation capability, that is, the ability to create rules on-line for the computer to use in identifying targets, significantly improved performance for the manual and override conditions [$\bar{X}_{O:NA}$=59.6, $\bar{X}_{O:A}$=69.7, SS_H=.15, SS_E=.102, $F(1,13)$=19.155, p=.001], but not the screening condition. In fact, override with allocation ($\bar{X}_{O:A}$=69.7) did not perform significantly worse than screening with allocation ($\bar{X}_{S:A}$=75.6); both performed significantly better than manual with allocation at the $p<.01$ significance level [e.g., $\bar{X}_{O:A}$=69.7, $\bar{X}_{M:A}$=57.4, SS_H=.221, SS_E=.307, $F(1,13)$=9.355, p=.009].

Table 3-20 presents the "mean percent correct" for the N=57 targets, with an initial degree of belief between .6 and .8.

Table 3-20

Experiment 2: The "Mean Percent Correct" for N=57 Targets with Initial Degree of Belief Between .6 and .8

		Allocation (i.e., Rule-Creation) Capability	
		No Allocation Capability	Allocation Capability
Human-Machine Interface Factor	Completely Manual	36.7*	66.2
	Override	72.1	77.1
	Screening	67.2	76.1

*Data from High-Workload Manual Condition in Experiment 1

The principal findings were as follows:

1. All five conditions significantly improved performance over that achieved in the manual:high-workload condition in Experiment 1. However, the screening without allocation and manual with allocation conditions in Experiment 2 did not improve performance over that achieved with the totally automated system (percent correct = 64.28).

2. The override:no allocation condition resulted in significantly better performance for targets with an initial degree of belief between .6 and .8 than did screening:no allocation [$\bar{X}_{O:NA}$=72.1, $\bar{X}_{S:NA}$=67.2, SS_H=.033, SS_E=.073, $F(1,13)$=5.974, p=.03], and approached significance for the manual:allocation condition [$\bar{X}_{O:NA}$=72.1, $\bar{X}_{M:A}$=66.2, SS_H=.049, SS_E=.153, $F(1,13)$=4.125, p=.063].

3. Having an allocation capability improved performance for all three interfaces. Even without a statistical test, there is a marked increase in performance for the manual:allocation condi-

tion over that achieved with the manual:no allocation condition under high workload. In addition, screening with an allocation capability resulted in a significant improvement in performance over that achieved without allocation [$\bar{X}_{S:A}=76.1$, $\bar{X}_{S:NA}=67.2$. $SS_H=.111$, $SS_E=.237$, $F(1,13)=6.073$, $p=.03$]. Although performance for the override condition improved with an allocation capability ($\bar{X}_{O:A}=77.1$, $\bar{X}_{O:NA}=72.1$), the increase was not significant ($p=.2$). Override with allocation did, however, result in significantly better performance than the manual:allocation condition [$\bar{X}_{O:A}=77.1$, $\bar{X}_{M:A}=66.2$. $SS_H=.166$, $SS_E=.190$, $F(1,13)=11.365$, $p=.005$], and the screening:no allocation condition [$\bar{X}_{O:A}=77.1$, $\bar{X}_{S:NA}=67.2$. $SS_H=.137$, $SS_E=.243$, $F(1,13)=7.346$, $p=.018$]. The increase in performance achieved with screening:allocation over that achieved with manual:allocation approached significance [$\bar{X}_{S:A}=76.1$, $\bar{X}_{M:A}=66.2$. $SS_H=.137$, $SS_E=.439$, $F(1,13)=4.067$, $p=.065$].

Table 3-21 presents the "mean percent correct" for the N=115 targets for which the initial degree of belief was <.80. The results of paired comparison tests replicated all the findings for the subset of N=58 targets with an initial degree of belief ≤.6, not the findings for the subset of N=57 targets with an initial degree of belief between .6 and .8. In particular, screening without allocation resulted in significantly better performance than override without allocation [$\bar{X}_{S:NA}=70.2$, $\bar{X}_{O:NA}=65.2$, $SS_H=.035$, $SS_E=.048$, $F(1,13)=9.513$, $p=.009$], and manual with allocation [$\bar{X}_{S:NA}=70.2$, $\bar{X}_{M:A}=61.1$, $SS_H=.115$, $SS_E=.19$, $F(1,13)=7.886$, $p=.015$] for the N=115 targets with an initial degree of belief <.8. The better performance without allocation achieved with screening than override is a function of the much larger mean difference favoring screening for targets with an initial degree of belief ≤.6 ($\bar{X}_{S:NA}=74.2$, $\bar{X}_{O:NA}=59.6$) than that favoring override for targets with an initial degree of belief between .6 and .80 ($\bar{X}_{S:NA}=67.2$, $\bar{X}_{O:NA}=72.1$). Again, the allocation capability had a marked effect for all three interfaces [$\bar{X}_{M:NA}=30.7$, $\bar{X}_{M:A}=61.1$; $\bar{X}_{O:NA}=65.2$, $\bar{X}_{O:A}=72.9$, $SS_H=.083$, $SS_E=.094$, $F(1,13)=11.539$, $p=.005$; $\bar{X}_{S:NA}=70.2$, $\bar{X}_{S:A}=75.2$, $SS_H=.035$, $SS_E=.098$, $F(1,13)=4.726$, $p=.049$], with the (now) statistical significance of the increase for screening being a function of the larger number of targets. Again, the performance levels achieved for override and screening with the allocation capability were statistically equivalent.

Table 3-21

Experiment 2: The "Mean Percent Correct" for N=115 Targets with Initial Degree of Belief Below .80

		Allocation (i.e., Rule-Creation) Capability	
		No Allocation Capability	Allocation Capability
Human-Machine	Completely Manual	30.7*	61.1
Interface	Override	65.2	72.9
Factor	Screening	70.2	75.2

*Data from High-Workload Manual Condition in Experiment 1

Table 3-22 presents the "mean percent correct" for the N=56 targets with an initial degree of belief <.80 that could be perfectly identified by messages. These data show the advantage of the rule-creation capability for messages that was available in the allocation condition. In particular, for the first and only time, manual with allocation resulted in better performance than that achieved with the screening without allocation and the override without allocation conditions, with the latter being statistically significant [$\bar{X}_{M:A}=75.1$, $\bar{X}_{O:NA}=63.4$, $SS_H=.192$, $SS_E=.367$, $F(1,13)=6.788$, $p=.022$]. Moreover, performance achieved in the manual:allocation condition was not significantly lower than that achieved in the override:allocation and screening:allocation condition. These results clearly demonstrate the value of giving the human operator some means of allocating work to the machine under high workload.

Table 3-22

Experiment 2: The "Mean Percent Correct" for N=56 "Message Targets"

		Allocation (i.e., Rule-Creation) Capability	
		No Allocation Capability	Allocation Capability
Human-Machine	Completely Manual	33.4*	75.1
Interface	Override	63.4	77.5
Factor	Screening	69.2	80.3

*Data from High-Workload Manual Condition in Experiment 1

Table 3-23 presents the "mean percent correct" for the N=85 targets with an initial degree of belief ≥.80. These results tend to replicate those achieved for targets with an initial degree of belief between .6 and .8. First, both the screening:no allocation and manual:allocation conditions did not improve performance over that achieved with the totally automated system (percent correct = 87); in fact, the manual:allocation condition resulted in poorer performance [$\bar{X}_{M:A}=75.1$, $SS_H=.105$, $SS_E=.298$, $F(1,13)=4.60$, $p=.05$]. The override:no allocation, override:allocation and screening:allocation conditions did result in significantly better performance than the totally automated system. Second, the override:no allocation condition resulted in significantly better performance than the screening:no allocation condition [$\bar{X}_{O:NA}=88.7$, $\bar{X}_{S:NA}=87.2$, $SS_H=.003$, $SS_E=.005$, $F(1,13)=7.822$, $p=.015$]. In addition, both nonallocation conditions resulted in significantly better performance than the manual:allocation condition [$\bar{X}_{O:NA}=88.7$, $\bar{X}_{M:A}=78.7$, $SS_H=.14$, $SS_E=.307$, $F(1,13)=5.92$, $p=.03$; $\bar{X}_{S:NA}=87.2$, $\bar{X}_{M:A}=78.7$, $SS_H=.103$, $SS_E=.30$, $F(1,13)=4.45$, $p=.055$]. Again, the allocation capability improved performance for all three interfaces. The increase for the manual interface is again extreme; in fact, it is stunning how poorly actual air defenders performed under high-workload in the first experiment for these "easy targets" without some (allocation) assistance from the machine. Although small, the increase for the screening condition is significant [$\bar{X}_{S:NA}=87.2$, $\bar{X}_{S:A}=88.6$, $SS_H=.003$, $SS_E=.002$, $F(1,13)=13.471$, $p=.003$]. Although the increased performance achieved for the override:allocation condition over that achieved for the override:nonallocation condition is not significant [$\bar{X}_{O:NA}=88.7$, $\bar{X}_{O:A}=89.7$,

SS_H=.002, SS_E=.008, F(1,13)=2.654, p=.127], it is significant over that achieved for the screening:allocation condition [$\bar{X}_{O:A}$=89.7, $\bar{X}_{S:A}$=88.6, SS_H=.002, SS_E=.003, F(1,13)=7.559, p=.017].

Table 3-23

Experiment 2: The "Mean Percent Correct" for N=85 Targets with Initial Degree of Belief ≥ .80

		Allocation (i.e., Rule-Creation) Capability	
		No Allocation Capability	Allocation Capability
Human-Machine Interface Factor	Completely Manual	36.1*	78.7
	Override	88.7	89.7
	Screening	87.2	88.6

*Data from High-Workload Manual Condition in Experiment 1

Table 3-24 presents the "mean percent correct" for the entire 200-target simulation.

Table 3-24

Experiment 2: The "Mean Percent Correct" for the Entire 200-Target Simulation

		Allocation (i.e., Rule-Creation) Capability	
		No Allocation Capability	Allocation Capability
Human-Machine Interface Factor	Completely Manual	33.0*	68.6
	Override	75.1	80.0
	Screening	77.4	80.9

*Data from High-Workload Manual Condition in Experiment 1

The principal results are as follows:

1. The manual interface with an allocation capability resulted in an extremely large increase in performance under high workload [$\bar{X}_{M:A}$=68.6, $\bar{X}_{M:NA}$=33.0]. The performance achieved with the manual:allocation interface was, however, not different than that achieved with a totally automated system using only the initially available information (percent correct = 69.0). Moreover, for the entire 200-target simulation, the mean performance achieved with the manual:allocation interface was significantly lower than that achieved for all four of the other interface conditions in Experiment 2 [e.g., $\bar{X}_{M:A}$=68.6, $\bar{X}_{O:NA}$=75.1, SS_H=.06, SS_E=.177, F(1,13)=4.44, p=.055]. These results suggest that giving the human operators initial target

identifications based on rules stored in the system better permits them to take advantage of additional information not available to the system and, thereby, improves performance over that achieved with only a "rule-creation" capability for allocating tasks (in our case, target identifications).

2. Supporting our hypothesis, screening without allocation resulted in significantly better performance than override without allocation [$\bar{X}_{S:NA}=77.4$, $\bar{X}_{O:NA}=75.1$, $SS_H=.007$, $SS_E=.017$, $F(1,13)=5.682$, $p=.033$]. This result was due to the substantially better performance (i.e., large mean difference) achieved with the screening:no allocation condition than the override:no allocation condition for the N=58 targets with an initial degree of belief $\leq .60$, for the latter condition actually resulted in small, but significantly better performance than the former for the N=57 and N=85 targets with initial degrees of belief between .6 and .80, and $\geq .80$, respectively. These results suggest that air defenders using a screening capability without allocation will indeed focus their attention on, and perform better for, the more difficult identification targets, but that there may be costs in terms of their ability to ensure the accurate identification of other (presumably easier) targets upon which their attention is not focused.

3. As hypothesized, the allocation capability significantly improved performance for all three interfaces [$\bar{X}_{M:A}=68.6$, $\bar{X}_{M:NA}=33.0$; $\bar{X}_{O:A}=80.0$, $\bar{X}_{O:NA}=75.1$, $SS_H=.034$, $SS_E=.033$, $F(1,13)=13.042$, $p=.003$; $\bar{X}_{S:A}=80.9$, $\bar{X}_{S:NA}=77.4$, $SS_H=.017$, $SS_E=.03$, $F(1,13)=7.386$, $p=.018$]. However, in contrast to our hypothesized rank order, override with allocation resulted in a higher mean level of performance than the screening without allocation condition, although the difference was not statistically significant [$\bar{X}_{O:A}=80.0$, $\bar{X}_{S:NA}=77.4$, $SS_H=.01$, $SS_E=.043$, $F(1,13)=2.863$, $p=.114$]. Moreover, screening with allocation did not result in significantly better performance than override with allocation [$\bar{X}_{S:A}=80.9$, $\bar{X}_{O:A}=80.0$, $SS_H=.001$, $SS_E=.055$, $F(1,13)=.242$, $p=.631$] for the entire 200-target simulation. We have generated three hypotheses for explaining this result. The first hypothesis is that the allocation capability essentially permits the operator to turn the override interface into a screening interface. If this were true, we would expect operators to use different information-processing strategies in the override:no allocation, override:allocation, and both screening conditions, and similar strategies in the override:allocation and both screening conditions. The second hypothesis is that this may be so, but the two different interfaces still foster somewhat different information-processing strategies. And the third hypothesis is that there are "ceiling effects," and mean performance can not exceed about .80 (or so). We can categorically assert that the last hypothesis is not true because 195 of the 200 targets (i.e., 97.5%) could be correctly identified on the basis of all the information.

We now turn to discuss the results of the information-processing analyses.

3.5.2.2 *Information-Processing Strategy.* The information-processing analyses focused on. (a) the percentage of targets hooked; (b) the length of hooking (in seconds); (c) the number of items of information requested; and (d) the percentage of times these items were requested in the five interface conditions in Experiment 2. In addition, we examined the types of rules generated in the three allocation conditions. Each of these analyses is considered, in turn.

Table 3-25 presents the mean values for "percentage of targets hooked by class" for each of the five conditions and three target classes, for the two override and two screening conditions, and for the two nonallocation and two allocation conditions for override and screening (i.e., not including the manual:allocation condition).

Table 3-25

Mean Percentage of Targets Hooked by Class

Human-Machine Condition	Degree of Belief Based on Initial Information			
	≤.6	.6 to .8	≥.80	$\bar{X}$
Manual + Allocation	51.4	42.6	29.2	41.1
Override	55.4	72.9	29.2	52.5
Screening	82.0	42.4	10.8	45.1
Override + Allocation	46.4	46.9	26.3	39.9
Screening + Allocation	58.4	32.4	08.7	33.2
Override (without and with Allocation)	51.1	61.4	27.8	46.8
Screening (without and with Allocation)	70.2	37.6	09.8	39.2
Nonallocation (O+S)	68.9	59.2	20.0	49.4
Allocation (O+S)	52.4	39.8	17.5	36.6

Paired comparisons showed that participants hooked a significantly higher percentage of targets with an initial degree of belief ≤ 60 in the screening than override condition [$\bar{X}_S$=70.2, $\bar{X}_O$=51.1, SS_H=2.054, SS_E=1.057, $F(1,13)$=25.253, p<.001]. In contrast, they hooked a significantly higher percentage of targets with degree of belief between .6 and .8 [$\bar{X}_O$=61.4, $\bar{X}_S$=37.6, SS_H=2.054, SS_E=1.057, $F(1,13)$=25.253, p<.001] and ≥.8 [$\bar{X}_O$=27.8, $\bar{X}_S$=9.8, SS_H=1.819, SS_E=.808, $F(1,13)$=29.268, p<.001] in the override than screening condition. This clearly demonstrates that the override and screening conditions affected air defenders' information-processing strategies and, in turn, performance for different classes of targets. It also affected their overall performance, for participants achieved higher mean performance values (significantly so without allocation) in the screening than override conditions, for both the N=115 test targets and the entire N=200 target simulation.

Paired comparisons also showed that when participants were in the override and screening conditions, they hooked significantly fewer targets with initial degree of belief $\leq.60$ and between .6 and .8 in the allocation than nonallocation condition [$\bar{X}_A=52.4$, $\bar{X}_{NA}=68.9$, $SS_H=1.514$, $SS_E=1.294$, $F(1,13)=15.205$, $p=.002$; $\bar{X}_A=39.8$, $\bar{X}_{NA}=59.2$, $SS_H=2.099$, $SS_E=1.071$, $F(1,13)=25.406$, $p<.001$]. Since their performance also improved in the allocation condition, the information-processing analyses suggest that, on the average, they were able to effectively create rules to deal with targets that they previously had to examine in the nonallocation condition. Interestingly, the mean values for the manual:allocation condition are comparable to the mean of the mean values for allocation with the override and screening conditions, except for considerably higher values for the manual condition for targets with an initial degree of belief $\geq.80$. Participants did not need to focus on these targets, particularly in the screening condition where they were classified as "firmly identified," as indicated by the lower mean values for screening in both the nonallocation and allocation conditions.

Figure 3-5 graphically presents the mean percentage of targets hooked by interface condition and initial degree of belief.

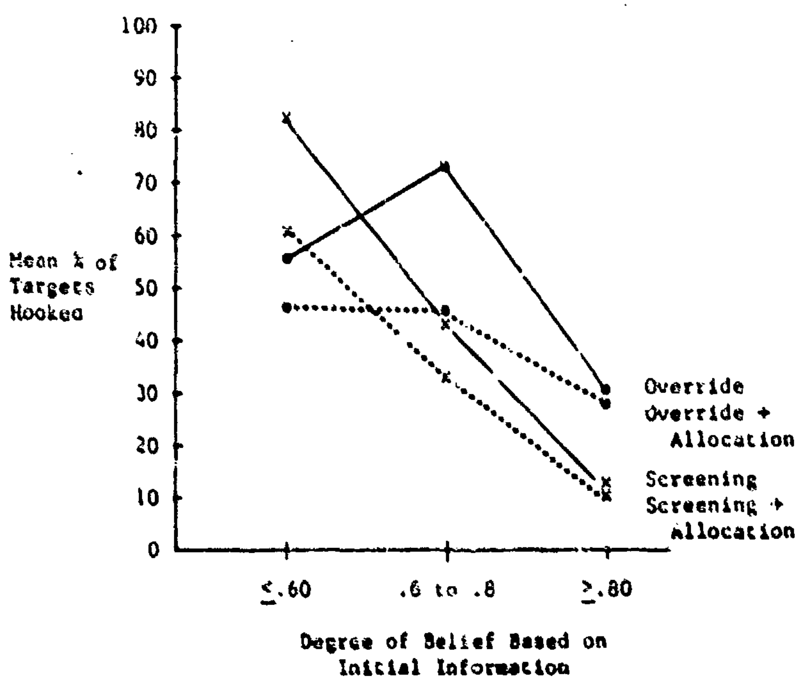


Figure 3-5. Mean percentage of targets hooked by interface condition and initial degree of belief.

The data for the override:allocation condition looks similar to that for the override:no allocation condition, except for targets with an initial degree of belief between .6 and .8. For this class of targets, the data

are more similar to that for the two screening conditions. In contrast, the data for the screening:allocation condition looks similar to that for the screening:no allocation condition except for targets with an initial degree of belief ≤.60, which is more similar to the two override conditions. These data suggest that the override and screening conditions foster different information-processing strategies. In addition, the allocation capability affects these different strategies by permitting operators to create rules to deal with targets they had to hook previously without an allocation capability.

Table 3-26 presents the mean values for "length of hooking by class" for each of the five conditions and three target classes, for the two override and two screening conditions, and for the two nonallocation and allocation conditions for override and screening. Examination of Table 3-26 shows that participants spent more time per hooking in the screening than override condition for all three classes of targets. Paired comparisons showed that significance levels were p=.049, p=.059, and p=.138 for the ≤.6, between .6 and .8, and ≥.8 classes, respectively; the difference for the last class did not approach significance because of a large mean squared error term. Paired comparisons also showed that participants spent significantly more time per hooking for the <.6 (p=.008) and between .6 and .8 (p=.04) targets when in the allocation condition. The mean value for the manual with allocation condition was comparable to that for the mean of the mean values for allocation with the override and screening conditions.

Table 3-26

Mean Length of Time (in Seconds) of Targets Hooked by Class

Human-Machine Condition	Degree of Belief Based on Initial Information <.6	.6 to .8	>.80	X
Manual + Allocation	13.79	13.49	12.92	13.40
Override	9.58	9.46	8.97	9.34
Screening	10.90	10.25	12.40	11.18
Override + Allocation	12.84	11.97	10.18	11.67
Screening + Allocation	14.49	15.37	20.63	16.83
Override (without and with Allocation)	11.21	10.71	9.58	10.50
Screening (without and with Allocation)	12.69	12.81	16.52	14.01
Nonallocation (O+S)	10.24	9.86	10.68	10.26
Allocation (O+S)	13.66	13.67	15.41	14.25

Figure 3-6 graphically presents the mean length of time data for the different interface conditions and initial degree-of-belief classes. Again, the reader is cautioned that the mean values between conditions were not significantly different for the targets with an initial degree of belief ≥.80 because of a large mean squared error term. Instead, the reader should focus on the form of the functions between interface conditions. Specifically, the functional form for override with allocation is more similar to that for override without allocation than the two screening

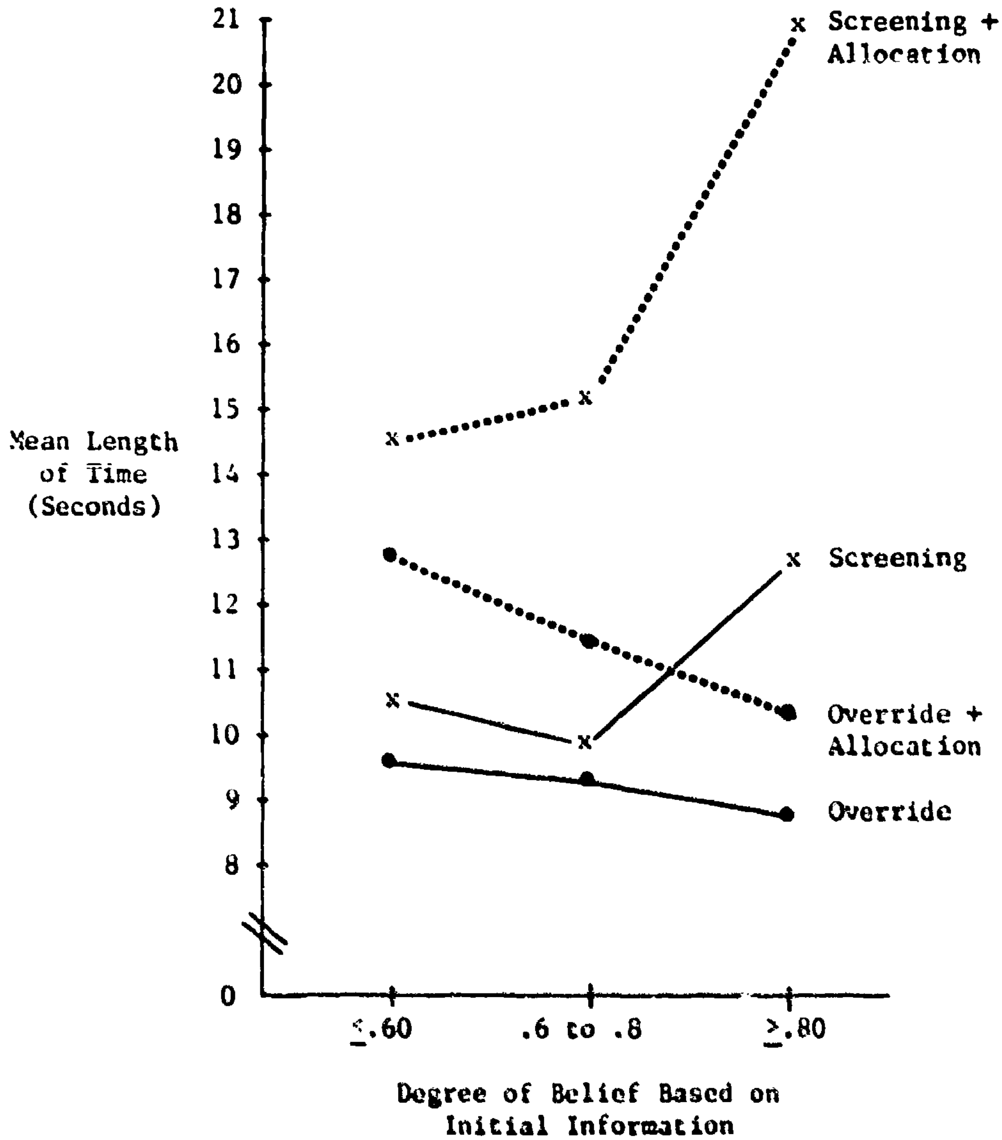


Figure 3-6. Mean length of time (in seconds) of targets hooked by interface condition and initial degree of belief.

conditions. Again, these data suggest that having an allocation capability did not turn the override interface into a screening interface; it just permitted participants to take somewhat longer to examine targets per hooking.

The third information-processing analysis was for the mean number of items of information requested per hooking with different interfaces. The mean number of items requested were override:no allocation = 2.05; screening:no allocation=2.682; manual:allocation=2.011; override:allocation=2.101; and screening:allocation=2.76. A repeated-measures ANOVA found a main effect for interface [SS_H=7.576, SS_E=27.304, $F(1,13)=3.607$, $p=.011$]. Examination of the means again shows a distinct difference between the override and screening conditions.

The fourth information-processing analysis focused on the percentage of times the participants using the different interfaces requested information about the four principal cues: IFF, HQ ID, corridor, and pop-up. These data are presented in Table 3-27, along with the row and column means of the data for comparison purposes. A repeated-measures ANOVA found a significant main effect for interface [SS_H=8447.6, SS_E=29,995.6, $F(4,52)=3.661$, $p=.011$] and for cue [SS_H=82,209.5, SS_E=101,924.2, $F(3,39)=10.485$, $p<.001$]. Consistent with the data for the mean number of items requested per hooking, there were larger mean percentages for the screening than override (and manual:allocation) conditions. Participants also requested IFF and HQ ID much more frequently than information about corridor and pop-up. Although IFF was more frequently requested than HQ ID when participants were in the override conditions, and less frequently in the screening conditions, the interaction was not statistically significant.

Table 3-27

Experiment 2: Percentage of Time Different Cues were Requested by Condition

	HQ	IFF	CORR	POPUP	$\bar{X}$
Override/NA	53.3	62.4	47.4	21.0	46.0
Screening/NA	71.1	63.5	54.9	23.7	53.3
Manual/A	48.4	61.9	38.3	11.4	40.0
Override/A	54.1	63.6	40.0	15.9	43.4
Screening/A	69.8	63.6	58.0	24.5	54.0
$\bar{X}$	59.3	63.0	47.7	19.3	47.3

In total, the information-processing analyses suggest that participants focused on different types of targets, took longer to examine them, and gathered more information about them when in the screening than override condition. With access to an allocation capability, participants also focused on different classes of targets and took longer to examine the targets that they chose to examine, letting the system use its prior rules and those generated by the participants to deal with more of the targets.

This "cooperative problem-solving" approach, under the allocation control of the human, resulted in the best overall performance.

Table 3-28 presents the types of rules created by three or more participants, as well as the number of instances the rule was created in total. The table is divided into two classes of rules: non-message and message-oriented rules. Within each class, the rules are ordered by the number of instances. Examination of the non-message rules indicates that participants primarily focused on the most diagnostic cue values (IFF +, Corridor:In, Corridor:Out, and Jammer:Yes), as indicated by our participating domain experts in Table 3-6 of this report. Interestingly, all participants did not create these rules in the manual:allocation condition; they, of course, did not have to create them with the override and screening interfaces. Moreover, some of the less frequently created rules have multiple clauses that appear unnecessary when compared with the more frequently created clauses. This suggests that more training in creating non-message-oriented rules when using the allocation condition, or perhaps a "knowledge-engineering" session toward this end with each participant, could have further improved performance. The latter possibility is particularly appealing because participants appeared quite capable of using the allocation capability to create message-oriented rules which explains why they performed so well in identifying the message-oriented targets with an initial degree of belief <.8. (The reader is reminded that "HQ ID" responses did not go into the system; consequently, no rules were created using this highly diagnostic cue.)

Table 3-28

The Types of Rules Created by Three (or More) Participants When Using the Allocation Capability

Type of Rule	# of Ps	# of Instances
Non-Message Rules		
IFF:+ → F	8	22
Corr:In → F	10	21
Corr:Out → H	8	18
Corr:Out and Jammer:Yes → H	7	15
Jammer:Yes → H	6	14
Corr:In and Jammer:No → F	4	7
Corr:Two-Out and Jammer:Yes → H	4	7
IFF:— → H	4	6
IFF:— and Jammer:Yes → H	4	6
IFF:+ and Corr:In → F	3	6
Corr:Two-Out → H	4	5
Message Rules		
Pop-up:FEBA, Corr:In, Alt:2000-5000, Sp:600-800 → H	13	30
Pop-up:FEBA, Corr:One-Out, Alt:2000-5000, Sp:1000+ → H	13	30
Pop-up:No, Corr:Two-Out, Alt:10000-80000, Sp:1000+ → H	13	29
Pop-up:Close, Corr:In, Alt:2000-5000, Sp:600-800 → H	12	28
Pop-up:FEBA, Corr:In, Alt:2000-5000, Sp:400-600 → H	3	4
Pop-up:Close, Corr:In, Alt:2000-5000, Sp:400-600 → H	3	3

Examination of the rules created, the information-processing analyses for the mean percentage of requests for different cues, and the subjective analysis for cue usage all indicate that participants considered certain cues more diagnostic than others. Moreover, although there were certainly individual differences, participants appeared to have a hierarchical (or ordered) sequence they went through to process cue information. Appendix E presents the cue-processing sequence that one of DSC's domain experts tended to follow in the nonallocation conditions for the manual, override, and screening interfaces. These results indicate that actual air defenders working with a realistic air defense simulation do not tend to use a "majority-of-confirming-dimensions" strategy to process information unless, of course, the values for different cues (i.e., dimensions) are basically contradictory and they are forced to count the cues "pro" and "con" the hypotheses.

3.5.2.3 *Workload*. The same two objective workload measures used in Experiment 1 were used in Experiment 2. Table 3-29 presents the means for all five conditions for both measures. Examination of the data in Table 3-29 suggests that adding the allocation capability does increase workload. The increase, however, did not tend to reach statistically significant levels. As in Experiment 1, we found no significant effects regarding the accuracy of the participants' responses to the "response light." The only paired comparison that approached significance was greater accuracy to the response light with screening:nonallocation than screening:allocation [$\bar{X}_{S:NA}$=89.07, $\bar{X}_{S:A}$=84.64, SS_H=274.57, SS_E=835.43, $F(1,13)$=4.27, p=.059]. Interestingly, the only statistically significant comparison for the second objective workload measure, mean response time to the response light, also found significant workload effects for adding the allocation capability to the screening condition [$\bar{X}_{S:NA}$=1179.1, $\bar{X}_{S:A}$=1324.6, SS_H=296092.6, SS_E=330341.4, $F(1,13)$=11.65, p=.005]. These (workload) results are consistent with the information-processing analysis showing that when participants were using the screening interface and the allocation capability, they gathered more data and took longer to examine targets than when they had only the screening interface. Although screening:allocation resulted in better performance than screening:nonallocation, one wonders if the performance difference might not have been greater except for the increased workload in the screening:allocation condition.

Table 3-29

Experiment 2: Mean Values for Both Objective Workload Measures for the Five Interfaces

	Mean Accuracy of Response to "Response Light"	Mean Speed (in msec.) of Response to "Response Light"
Override:No Allocation	89.21	1196.6
Screening:No Allocation	89.07	1179.1
Manual:Allocation	85.71	1283.9
Override:Allocation	87.57	1241.9
Screening:Allocation	84.64	1324.6

3.5.2.4 *Subjective Measures.* The same two questionnaires used in Experiment 1 to obtain participants' subjective performance, workload, and preference measures were used in Experiment 2. The results for each questionnaire are considered, in turn.

The first questionnaire was given immediately after the completion of each condition in the design. Participants were asked to rate, on a 9-point scale, their performance, the level of workload, and the degree to which they liked working with the system. Table 3-30 presents the mean responses for subjective performance, workload, and preference for each of the five conditions in the design. Higher values represent better scores on all three subjective measures.

Table 3-30

Experiment 2: Mean Subjective Performance, Workload, and Preference (First Questionnaire)

(a) *Subjective Performance*

		Allocation Capability	
		No	Yes
Human-Machine Interface Factor	Manual	3.73*	5.36
	Override	5.50	6.93
	Screening	5.71	6.57

(b) *Subjective Workload*

		Allocation Capability	
		No	Yes
Human-Machine Interface Factor	Manual	3.47*	4.14
	Override	4.50	6.07
	Screening	5.57	5.43

(c) *Subjective Preference*

		Allocation Capability	
		No	Yes
Human-Machine Interface Factor	Manual	4.67*	4.79
	Override	4.5	6.64
	Screening	5.2	6.14

*Data from high workload:manual condition in Experiment 1

Examination of Table 3-30 indicates that the allocation capability had a marked effect on the mean subjective performance rating for all three interfaces, a result that agrees fully with the objective performance data. A one-way, repeated-measures ANOVA on the subjective performance ratings for the five interface conditions in Experiment 2 was significant [SS_H=27.057, SS_E=80.943, $F(4,52)$=4.346, p=.004]. Paired comparisons showed that the mean subjective performance for the override:allocation condition was not significantly different than that for the screening:allocation

condition, but that both were significantly higher than that for all other conditions.

In addition to having the best mean subjective performance rating, the override:allocation condition had the best mean subjective workload rating. In particular, paired comparisons showed that, in contrast with the objective workload measures, the mean subjective workload rating for the override:allocation condition was significantly better than that for both the manual:allocation condition [$\bar{X}_{O:A}$=6.07, $\bar{X}_{M:A}$=4.14, SS_H=52.07, SS_E=80.93, F(1,13)=8.365, p=.013] and override:no allocation condition [$\bar{X}_{O:A}$=6.07, $\bar{X}_{O:NA}$=4.50, SS_H=34.574, SS_E=33.43, F(1,13)=13.444, p=.003]. Other significant paired comparisons included better mean subjective workload in the screening:no allocation condition than the manual:allocation condition [$\bar{X}_{S:NA}$=5.57, $\bar{X}_{M:A}$=4.14, SS_H=28.57, SS_E=67.43, F(1,13)=5.508, p=.035] and override:no allocation condition [$\bar{X}_{S:NA}$=5.57, $\bar{X}_{O:NA}$=4.50, SS_H=16.07, SS_E=30.93, F(1,13)=6.755, p=.022]. However, in contrast with the objective workload results, participants did not consider the screening interface to be more difficult with the allocation capability.

Examination of the subjective preference data again shows the positive effect for the allocation capability, although the effect for the manual interface is small. The mean subjective preference rating for the override:allocation condition was significantly higher than that for all conditions except the screening:allocation condition. The mean subjective preference rating for the screening:allocation condition was significantly higher than that for the override:no allocation condition [$\bar{X}_{S:A}$=6.14, $\bar{X}_{O:NA}$=4.5, SS_H=37.79, SS_E=57.21, F(1,13)=8.586, p=.012], and almost significantly higher (i.e., p<.10) than the manual:allocation and screening:no allocation condition.

As in Experiment 1, the second questionnaire was given at the end of the experimental session in Experiment 2. Again, participants were asked to rate on a 9-point scale how well they think they performed with each of the systems, how hard they worked to perform the aircraft-identification task with each system, and how much they liked working with each system. Table 3-31 presents the mean responses for the three measures for each of the five conditions; again, higher values represent better scores.

Table 3-31

Experiment 2: Mean Subjective Performance, Workload, and Preference (Second Questionnaire; Higher Values Represent Better Scores)

	Performance	Workload	Preference
Override:No Allocation	5.29	3.93	4.79
Screening:No Allocation	6.29	5.14	5.50
Manual:Allocation	5.50	4.79	5.43
Override:Allocation	6.64	5.14	5.57
Screening:Allocation	6.50	5.64	6.43

(The mean values for the manual condition in Experiment 1 are not included in Table 3-31 because the second questionnaire in Experiment 1 did not distinguish between high and low workload; consequently, the values are not comparable to those in Experiment 2, which are only for high workload.)

As with the first subjective questionnaire and the objective performance data, the override:allocation and screening:allocation conditions had the two highest mean subjective performance ratings. Although the mean subjective performance rating for the override:allocation condition was not significantly higher than that for the screening:no allocation condition, it was significantly higher than that for the override:no allocation condition [$\bar{X}_{O:A}$=6.64, $\bar{X}_{O:NA}$=5.29, SS_H=25.79, SS_E=27.21, $F(1,13)$=12.32, p=.004] and the manual:allocation condition [$\bar{X}_{O:A}$=6.64, $\bar{X}_{M:A}$=5.5, SS_H=18.286, SS_E=37.714, $F(1,13)$=6.303, p=.026]. The mean subjective performance rating for the screening:allocation condition was significantly higher only than that for the override:no allocation condition [$\bar{X}_{S:A}$=6.5. $\bar{X}_{O:NA}$=5.29, SS_H=20.643, SS_E=40.357, $F(1,13)$=6.65, p=.023].

The screening:allocation, override:allocation, and screening:no allocation conditions also had the best mean subjective workload scores. In contrast with the first questionnaire, however, none of the paired comparisons were statistically significant at the $p<.05$ level. This occurred because of considerable variability in the subjective workload measures for the second questionnaire and, consequently, large SS_E terms.

There also was large variability in the mean subjective preference ratings. The only significant paired comparison was the increased preference for the screening interface with an allocation capability [$\bar{X}_{S:A}$=6.43, $\bar{X}_{S:NA}$=5.50, SS_H=12.071, SS_E=32.929, $F(1,13)$=4.766, p=.048].

The second questionnaire also asked participants to rate how much each cue affected their ability to obtain correct identification, thus providing a subjective measure of cue utility. The mean cue utility measures are shown in Table 3-32.

Table 3-32

Experiment 2: Mean Utility Rating for the Target Identification Cues

Cue	Rating	Cluster
IFF	7.714	Cluster 1
HQ ID	7.571	Cluster 1
Corridor	6.714	Cluster 1
Jamming	6.571	Cluster 1
Messages	5.290	Cluster 1
Speed	4.286	Cluster 2
Altitude	3.571	Cluster 2
Heading	3.286	Cluster 2
Pop-Up	3.214	Cluster 2
Distance	2.929	Cluster 2

Paired-comparison tests suggest that the cues can be grouped into roughly two clusters. Specifically, the first cluster includes IFF, HQ ID, Corridor, Jamming, and Messages; among these cues, the only significant difference was between IFF and Messages [$\bar{X}_I=7.714$, $\bar{X}_M=5.29$, $SS_H=44.643$, $SS_E=88.357$, $F(1,13)=6.568$, $p=.024$]. The second cluster includes Speed, Altitude, Heading, Pop-Up and Distance; among these cues, the only significant difference was between Speed and Distance [$\bar{X}_S=4.286$, $\bar{X}_D=2.929$, $SS_H=25.786$, $SS_E=45.214$, $F(1,13)=7.414$, $p=.017$]. All of the cues in the first cluster had significantly higher mean usefulness ratings than each cue in the second cluster.

The mean usefulness ratings obtained in Experiment 2 are comparable to those obtained in Experiment 1. In both experiments, IFF, HQ ID, and Corridor had the three highest mean ratings and Heading, Pop-Up, and Distance had three lowest mean ratings. Jamming and Messages had higher mean ratings, and Speed and Altitude had lower mean ratings in Experiment 2.

In addition, the rank-order of mean subjective ratings for IFF, HQ ID, Corridor, and Pop-Up matches that obtained in the information-processing analysis. In the latter, however, IFF and HQ ID each had significantly higher mean values than Corridor.

We now turn to a discussion of the results of the Fort Bliss experiments in Section 4.0 of this report.

4.0 CONCLUSIONS

Our research investigating the relative effectiveness of alternative human-machine allocation schemes has, in large part, been guided by two information-processing principles resulting from cognitive psychological research under high-workload conditions:

(1) Decision makers should focus their attention on those cases that require their attention. Other cases should be delegated to others (people or machines) who can solve them.

(2) Given that people are focusing on the cases needing their attention, they need to use an appropriate information-processing strategy to solve these cases.

These principles were supported in the Phase I research; under high workload, only the "screening" condition made participants primarily attend to the unresolved targets and, to a lesser extent, more heavily rely on the "extra cue." These principles were supported in the Phase II research, particularly in the second Fort Bliss experiment; under high workload, the best performance was achieved by adding an allocation (i.e., rule-creation) capability to the screening and override interfaces, respectively, thereby permitting participants to gather more information and take longer to examine those targets requiring their attention. Both phases of our research demonstrated that when successfully implemented, cooperative problem-solving approaches between the person and the machine were significantly more effective than the "manual only" or "totally automated" approaches to solving complex inference tasks.

The above "basic principles" have been one component guiding our efforts to expand the "screening" condition in Phase I into more collaborative human-machine allocation schemes for investigation in Phase II. A second component has been the development of a more fundamental theoretical understanding of the psychological mechanisms underlying human-computer performance. This second thrust makes possible a more integrative, more generalizable, and more quantitative design technology than accumulating isolated, albeit powerful, principles by permitting one to model information-processing strategies. By successfully linking models of information-processing strategies to performance with different human-machine interfaces, we have moved toward the long-term goal, of which the current research program is but a small part, of developing cognitive human-factors technology that is a theory-based, quantitative methodology for predicting human-computer system performance under diverse conditions

In presenting the results of our experiments in Sections 2.0 and 3.0 of this report, we have stayed, so to speak, very close to our data. In closing this report, however, we attempt to generalize from our controlled experiments to the real-world Army environment. We, of course, do so with some trepidation and urgings of caution, for additional experimental research directed toward developing models of information-processing strategies is required before achieving our long-term goal of developing

cognitive human-factors technology. We fully appreciate, however, the need to relate our findings to ongoing requirements of personnel designing person-machine interfaces in air defense and other Army domains. Therefore, Sections 4.1 and 4.2 of the report discuss the results of the experiments from the perspective of (a) developing guidelines for human-machine interfaces and (b) extending both the findings and broader theoretical and methodological approach to other Army domains, respectively.

4.1 Toward Guidelines for Human-Computer Interfaces

The theoretically-based research performed to date supports the following guidelines for constructing human-computer interfaces for performing tasks involving the identification of a large number of objects (e.g., aircraft) under conditions of (a) high workload and (b) high uncertainty, because all the data are not resident in the computer's database.

1. Design the interface to support cooperative problem solving between the human and the machine. Do not develop "manual only" or "totally automated" systems.

2. Embed "identification" rules in the computer to give it the capability to solve "easy" problems not requiring the operator's attention. If the workload is excessively high, then implement some type of screening mechanism to bring "difficult" problems to the operator's attention. It is important to point out that both objective and subjective workload measures indicated that workload was not excessively high with the override interface in the two Fort Bliss experiments, whereas it was too high in the Phase I experiment. Therefore, the workload required by different interfaces needs to be considered from a cognitive perspective.

3. If possible, develop a flexible interface that permits the operator to develop rules for allocating (cognitive) tasks to the computer because this significantly improves performance. It is important to note that this capability was used extensively in the second Fort Bliss experiment prior to initiation of the problem runs, but not often once aircraft began approaching the air defense sector. Again, the workload level for this interface needs to be seriously considered if this interface is to be used "on-line."

4. The interface designer needs to give serious consideration to task characteristics. Our research indicates that interfaces (such as the "screening" interface) that are designed to focus attention on certain types of tasks (e.g., targets with initial diagnosticity <.6) may cause operators to examine more information about those tasks and thus take longer to solve them, perhaps to the detriment of other presumably easier tasks (i.e., targets with initial diagnosticity $\geq .6$). Therefore, the designer needs to give serious consideration to the characteristics of the task environment (e.g., the distribution of

targets with various initial diagnosticity probabilities). Through pilot-testing efforts, the designer can develop information-processing models for different types of interfaces, and via simulations, predict the anticipated performance levels for different, but expected, distributions of task characteristics.

5. Designers need to consider procedures for reducing the cognitive workload and speeding-up the information-processing time associated, on the average, with both the screening and the allocation capabilities. These capabilities resulted in the best performance in the second Fort Bliss experiment, but it appears that operators also expanded the task to the time available. Since even better performance was attainable in our task, faster information-processing with the screening and allocation capabilities may have resulted in even better performance with these interfaces.

6. The designer should assume that expert operators have a hierarchical (or ordered) sequence of cues (based on cue diagnosticity) that they go through to process information. The experiments with actual air defenders suggest that operators do not use a "majority-of-confirming-dimensions" strategy to process information unless, of course, (a) the values of different cues are contradictory and they are forced to count "pros and cons," or (b) the environment is such that the cues need to be equally weighted to achieve the best prediction. Again, all of this emphasizes the importance of understanding and modeling the human's information-processing strategy in order to best design the human-computer interface.

4.2 Implications for Army Command and Control

4.2.1 General applications. The concept for an integrated Army Command and Control System (ACCS), previously called SIGMA-STAR, will link five battlefield functional areas: maneuver control, fire support, air defense, intelligence and electronic warfare (IEW), and combat service support. The Army is presently engaged in a significant upgrade of the system, with the intent that a single system would be procured for each of the above five elements. Each of these systems will provide for communication and data-sharing among the various nodes within each of these elements, as well as some data-sharing and communication between elements. Each of these systems, in effect, represents a distributed database capability with manager (commander) interface support. Systems such as the Mobile Subscriber Equipment (MSE) will link the mission areas with secure jam-resistant communications provided by a comprehensive C^3 network of mobile radios, switching centers, and telephones.

The results of the Phase II experiments in the air defense mission area strongly suggest further experimental applications in the other four areas. In the area of maneuver control, a commander must effectively control his tactical elements, he must know where they are located and must

have a means by which to talk to them even in an enemy electronic countermeasure environment. ACCS will permit battlefield commanders to receive real-time combat data such as troop maneuvers and general battlefield conditions. New capabilities, such as the Enhanced Position Location Reporting System (EPLRS), the Joint Tactical Information Distribution System (JTIDS), and the Maneuver Control System (MCS), are currently being developed and fielded. Additional new capabilities may be provided by concepts such as remotely piloted vehicles (RPVs) and robotics (e.g., the Teleoperated Mobile Anti-Armor Platform (TMAP)) which will offer the commander new opportunities for data gathering, analysis, and even tactical operations. The increased information available to the tactical battle manager, much of it time-sensitive, has the potential to overwhelm his decision-making capability. This information includes friendly unit location updates down to platoon-sized elements, terrain and other battlefield conditions, and reported enemy positions. Maneuver control includes both real-time decision making on the maneuver activities of ground forces as well as longer-term considerations that may range from hours to days. A maneuver control decision aid which allows the commander to operate in the screening or override modes may improve the quality of decisions.

Fire support applications are very similar to the air defense problem. Fire support involves use of surface-to-surface missiles, artillery, and close-in support resources against enemy forces in support of maneuver units. Fire planning involves assessing enemy activities, predicting target types and locations, and allocating forces against the highest payoff targets. Larger numbers of ground targets will be available to the fire support element because of increasing target acquisition capabilities, including Joint Surveillance and Target Attack Radar System (JSTARS), Advanced Field Artillery Tactical Data System (AFATDS), Quick Look, Guardrail, and Position Locating Strike System (PLSS). As in the air defense application, tentative target identifications are made using multiple sources and "cues" The decision maker must select target priorities, assign weapon systems for interdiction (target distribution) and coordinate airspace and ground space.

Intelligence and electronic warfare (IEW) applications are also possible. The Joint Tactical Fusion Program is developing an all-source, real-time intelligence capability for the commander, but the advantages of such a system may be lost without a complementary decision support system. In particular, the development of order of battle and the intelligence preparation of the battlefield could be automated. The Order of Battle technician is particularly susceptible to make errors in judgment as information is coming in faster than he can evaluate it.

Finally, the combat service support mission area holds applications possibilities. The new Tactical Army Combat Service Support Computer System (TACCS) will support personnel functions in battalions and the maintenance and supply operations of direct support units. Systems such as the Logistics Application of Automated Marking and Reading Symbols (LOGMARS) offer the opportunity for real-time equipment management, and will feed directly into TACCS which will allow the Army to operate from a single database during peace and war. However, decisions about difficult systems [illegible] such as inventory control, shipping and routing, personnel re-

placement requisitioning and allocation of medical resources must still be made by staff officers supported by such systems. An automated system which can help make these decisions subject to "allocation" rules as well as to screening or override assistance to the human operator may improve the timeliness and quality of the decisions.

4.2.2 Additional testbed application. As indicated above, there are potential applications for the results of Phase II in each of the other battlefield functional areas. Based upon the criticality of need, state-of-the-art in fielded systems, and projections for new equipment, the intelligence/EW area appears to be the most appropriate choice for an additional testbed for the findings of this effort.

Intelligence support of the command and control (C^2) function is a key ingredient in achieving a decisive edge on the battlefield. The services have the need for an intelligence processing system that can collect all available intelligence, mold it into a total look at the battlefield, and present it to intelligence officers and commanders to support decisions rapidly. To that end, the Army and Air Force have initiated a program called Joint Tactical Fusion (JTF) that will allow the commander to seize the initiative by better understanding enemy intentions. The problem of tactical fusion is to transform a continuous stream of literally thousands of collection reports into a list of military units and their locations.

The major components of JTF are the Air Force's Enemy Situation Correlation Element (ENSCE) and the Army's All Source Analysis System (ASAS). Together, these systems operating as JTF will form the focal point for conduct of the AirLand Battle. The complete system is expected to be available in 3-4 years, but some components are already being readied. Software programs should be completed and ready for testing in 1988.

JTF is not designed to process raw data, but rather to accept as standard messages, information from various collection and analysis agencies. It is capable of accepting reports from any of the tactical or national systems such as Joint-Stars, NSA satellites or remote sensors. JTF will be able to integrate and correlate intelligence from communications, human, imagery, or emissions sources and develop a coherent estimate of enemy potential. It will also contain organizational, situational, and doctrinal templates for enemy forces that will allow analysts to estimate enemy capabilities and intentions based upon past performance.

The role of the human analyst in JTF remains critical since JTF is not designed to replace the intelligence analyst or the commander. At present, intelligence staffs are overloaded with information, particularly during periods of heavy combat; deal with multiple inputs from many sophisticated sources; are under considerable time stress as the battlefield status changes rapidly; and are involved in high-stakes tasks since the commander's decisions rely heavily upon intelligence input. Recognizing that these are the same characteristics that made the air defense functional area an appropriate choice for the Phase I and Phase II testbed for the effort, we suggest that the JTF arena is an appropriate testbed for further testing and validating of our conclusions.

At present, intelligence staffs are overwhelmed by manual operations. Reports must be sorted manually, analysis has little automation support, and there is only a modest degree of standardization across reports. Databases are extensive, and searching for key pieces of information is time-consuming, difficult, and often redundant. Multiple sources often provide conflicting or overlapping information, and correlation is a difficult task. Under JTF, many of these problems will be solved, but based upon our findings in Phase II, we believe that it would be more efficient if JTF included some of the guidelines for human-computer interfaces discussed in Section 4.1.

Of particular interest within the intelligence area is the order of battle (OB) analyst. Order of battle analysis produces an assessment of the identification, strength, command structure and disposition of the personnel, units, and equipment of the enemy military force. This is typically expressed to the commander in terms of an analysis of enemy probable courses of action based on collated order-of-battle information. The OB analyst receives information from a variety of sources and has access to numerous intelligence databases. He makes use of doctrinal templates (how does the enemy like to fight if unrestricted by terrain), situational templates (terrain constraints are applied), event templates (conditions and events to assist in collection planning), and decision-support templates (to match intelligence needs with tactical decision points). He must consider the following OB factors: composition, training, disposition, logistics, strength, combat efficiency, tactics and miscellaneous data. The OB analysts must deal with extensive reference material to include:

OB handbooks - background data on political structure, tactical doctrine, and military organization;

OB books - compilations of current intelligence which present composition and disposition of forces;

installation handbooks - detailed information on all military installations;

miscellaneous references.

The OB analyst is responsible for preparing and maintaining the OB workbook, situation maps, card files, personality files, strength worksheets, and for evaluation and interpretation of all OB elements.

Under the JTF program, the OB analyst will be greatly supported by automated systems. But many issues must be resolved that relate directly to the Phase II research results:

1. The interface should be able to support cooperative problem solving that does not rely solely on the human or solely on the automated system. For example, in a fully manual state, the OB analyst would receive unprocessed data from numerous sources (e.g., from corps HQ's report that enemy heavy bridging is moving forward in the vicinity of grid coordinates MB8634; from

surveillance satellites, nuclear-capable SCUD units have been sighted within 20 km of FEBA). He must manually examine each piece of data, determine its accuracy and the reliability of the source, and infer the implications of the data. For a fully automated mode, all data is directly inputted to the computer which processes and analyzes the information and draws inferences. Both the human and the computer have qualities that the other lacks. The human can visually correlate intelligence reports to terrain features or maps in a way that the computer cannot. The computer, however, can sort its way through masses of intelligence reports that the OB analyst can not possibly complete. A cooperative system could merge the advantages of both.

2. Rules need to be embedded in the system to allow it to solve problems not requiring the OB analyst's attention. These rules can be based on the templates described above and can be extended to include situational issues. For example, if the goal of the situation assessor (with the help of the OB analyst) is to identify and locate enemy units on a map as a function of intelligence reports, a prototypical system with an override capability could perform as follows:

   Intelligence summary reports (INTSUMs) are received as shown below:

   1. Enemy anti-tank gun sited in dug in position in vicinity of MB 8672.

   2. Combat patrol engages threat platoon in hasty defensive position, vicinity of MB 8651. The position had incomplete overhead cover, and communications trenches were being built.

   3. Enemy patrol sited vicinity MB 8942 emplacing mines by hand.

   The OB analyst would attempt to use this information to identify type and location of enemy units. He then would work in conjunction with the situation assessment analysts to determine enemy capabilities and intentions. Together they could use the computerized system to assist the intelligence analysis process.

   The computer would have a set of embedded rules and templates for various organizations, doctrine, and tactics. For example, a template for a deliberate defense might include:

   Deliberate defense:

   - is used to retain terrain;
   - makes heavy use of anti-tank guns in dug-in positions;

- is structured to include a security zone and defensive belts;
- makes extensive use of mines placed by hand and overhead coverage.

The embedded rules would include an inference system such as the Shaferian models used in Phase II, a Bayesian structure, or some other approach to quantifying uncertainty. This would enable the computer to look at the INTSUMs and derive a conclusion such as there is an 85% chance that the enemy intention is to prepare a deliberate defense.

In many cases, the indicators are clear and the resulting inferences have a high degree of credibility. However, in others there are conflicting reports making the picture less clear. For example, if the INTSUM report referred to above also indicated that heavy bridging was moving forward (an indicator of attack), there would be conflict in assessing enemy intentions. In this type of situation, the computer could use a screening condition to alert the OB analyst that further assessment is needed.

3. The interface of the OB analyst and his portion of JTF should be flexible enough to allow the OB analyst to develop rules for allocating tasks to JTF. The designer of the interface must carefully focus on the information-processing models used by the OB analyst on a variety of tasks. Some OB analysts focus on movement of artillery as the major indicator of enemy intent. Others look more closely at the actions of the combat service support elements as being more diagnostic. Still others will prefer to focus on the maneuver units. The results of Phase II indicate that an interface is needed that allows the individual analyst to tailor the way the system operates to his own cognitive style. The interface might allow for user preferences in knowledge representation to include scripts, frames, mental models, and production rules. For example, the user might specify the production rule:

IF	high activity in rear area and high activity is newly detected, and other units not moving rearward,
THEN	determine that new unit has entered rear area.

The system then correlates new intelligence inputs with the embedded knowledge base. Since the diagnosticity of the various indicators may be situation-dependent, provision should be made for the user to modify the quantitative factors in the model as appropriate. Unlike the air defense situation, the factors in this case would probably not change more frequently than once every 24 hours.

In attempting to apply the findings of Phase II to the OB analyst and JTF, a simulation on the IBM-PC could be developed that represents the OB analyst/JTF link. Since JTF is not fully fielded, its projected major characteristics could be emulated to included analysis capabilities, database access, and presentation of output. A prototype of the JTF is scheduled for user evaluation in December 1987 by the 552nd Military Intelligence Battalion at Fort Hood, Texas, and that prototype could serve as a model for the testbed. If the evaluation is successful it might be feasible to use the prototype itself as an experimental vehicle.

The research would address the same issues for the OB analyst that were addressed in the air defense testbed.

> *What is the impact of workload on human processing of information?* The OB analysts would have to perform their tasks under varying conditions of time stress, workload (in terms of pieces of information needing processing and degrees of conflict of information).
>
> *What is the impact of allocation schemes that represent cooperative problem solving?* The system could be made to operate in (1) a manual mode in which the computer serves only as a link to pass data; (2) an override mode in which the computer processes data to make OB assessments, presents them to the OB analyst, and allows him to override these assessments, and (3) a screening mode in which the computer brings to the attention of the OB analyst conflict situations regarding enemy intentions/capabilities.
>
> *How do flexible allocation schemes compare with fixed ones?* In the fixed schemes, the OB analyst and computer each has pre-set roles for allocation of tasks. In the flexible scheme, the OB analyst may vary the computer contribution as a function of the situational variables.

We would anticipate conducting controlled experiments regarding the relative effectiveness of alternative human-machine allocation schemes with actual U.S. Army OB analysts. This could be done in conjunction with the JTF user test at Fort Hood, or at Fort Huachuca at the intelligence school. We anticipate that the results of the psychological research would be useful in the final design of the human-machine interface for the OB portion of the JTF.

REFERENCES

Chinnis, J.O., Jr., Cohen, M.S., and Bresnick, T.A. (1985). *Human and computer task allocation in air-defense systems* (Army Research Institute Technical Report 691). Alexandria, VA: Army Research Institute.

Cohen, M.S., Bromage, R.C., Chinnis, J.O., Jr., Payne, J.W., and Ulvila, J.W. (1982). *A personalized and prescriptive attack planning decision aid* (Technical Report 82-4). Falls Church, VA: Decision Science Consortium, Inc.

Cohen, M.S. Brown, R.V., and Chinnis, J.O., Jr. (1986). *Tactical decision aiding for submarine attack planning* (Technical Report 86-2). Falls Church, VA: Decision Science Consortium, Inc.

Cohen, M.S., Watson, S.R., and Barrett, E. (1985). *Alternative theories of inference in expert systems for image analysis* (Technical Report 85-1). Falls Church, VA: Decision Science Consortium, Inc.

Feller, W. (1950). *An introduction to probability theory and its applications*. NY: Wiley and Sons.

Payne, J.W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance, 16*, 366-387.

Payne, J.W., Bettman, J.R., and Johnson, E.R. (1986). *Adaptive strategy selection in decision making* (Technical Report 86-1). Durham, NC: Duke University.

Phelps, R.H., Halpin, S.M., and Johnson, E.M. (1981). *A decision support framework for decision aid designers* (Technical Report 504). Alexandria, VA: U.S. Army Research Institute.

Rouse, W.B. (1977). Human-computer interaction in multitask situations. *IEEE Transactions on Systems, Man, and Cybernetics, 7*(5), 384-392.

Shafer, G. (1976) *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.

Shafer, G. (1981). Jeffrey's rule of conditioning. *Phil. of Sci., 48*, 337-362.

Shafer, G., and Tversky, A. (1983). *Weighing evidence: The design and comparison of probability thought experiments*. Stanford, CA: Stanford University.

Singleton, W.T. (1974). *Man-machine systems*. Baltimore, MD. Penguin Books, Inc.

APPENDIX A

INSTRUCTIONS DESCRIBING THE PARTICIPANTS' "BASIC JOB" AND THE AIR DEFENSE TESTBED

A.1 Purpose of Study

The study you are taking part in is designed to help the Army learn how to design future air defense systems so that they efficiently allocate tasks between the person and the computer under different levels of workload.

A.2 Your Basic Job

You will be the tactical control officer (TCO) of a proposed air defense system and must decide which of the aircraft approaching you are friends and which are hostiles. You will see a "radar" picture of the sector you are responsible for defending, and--on another portion of the screen--you will find a variety of information that, in addition to the radar display, will help you make identification decisions. Your job is to accurately identify incoming aircraft so that as many enemy aircraft as possible are shot down and that as few friendly aircraft as possible are shot down. Do not think in terms of "weapons tight," "weapons hold," or "weapons free." Instead, think only of "correct identification."

Figure A-1 shows what the radar screen looks like. You are located at the bottom, where the two straight lines come together. The whole pie-slice-shaped area is the area your radar can see. You are protecting two friendly assets. In addition, there are two safe-passage corridors for the movement of friendly aircraft. These corridors are outlined in blue.

The "U" shaped figures on the screen represent the location of UNKNOWN aircraft. These aircraft have not been identified as either friends or hostiles. Aircraft will fly at different speeds and altitudes depending on whether they are bombers, fighters, or helicopters. Aircraft may appear at the extreme top or sides of the screen or they may "pop-up" within the sector if they had been flying at an altitude below radar detection. If they pop-up, they may do so "close" to you or along the FEBA. Aircraft may be outside or inside a safe-passage corridor. Aircraft that are close to the edge of the corridor are classified as "inside" if they trace the edge of the corridor. If inside a corridor, they may be flying within the altitude and speed parameters set for the corridor or not. The acceptable corridor altitude ranges from 2,000 to 10,000 feet; the acceptable corridor speed is 700 knots or below. In all cases, the aircraft will move either toward you or the sides of the sector; never toward its top. You may use pop-up criteria, together with your other ID means, to determine if an aircraft is friendly or hostile. Unfortunately, there has been heavy enemy jamming in the area of the corridor entrances and it has been impossible to use the corridor entry point as a reliable ID criterion. Both friendly and

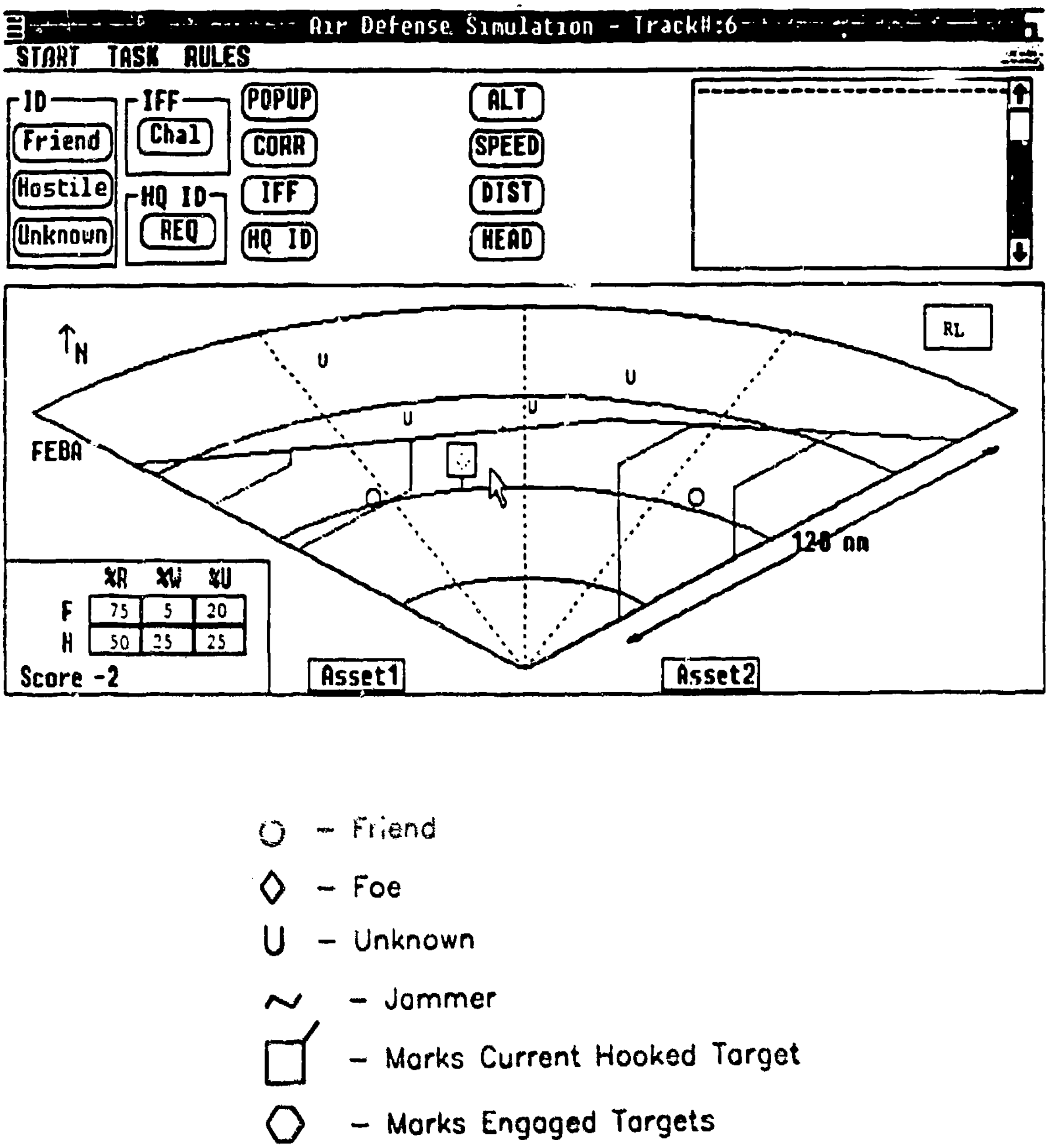


Figure A-1. The radar screen and basic symbols.

hostile aircraft have been appearing for the first time well within the corridor.

To perform well, you must correctly identify the type of as many aircraft (i.e., friend or hostile) as possible before they go off the radar screen. Aircraft will go off the screen when they have reached the sides of the sector or when they are 40 km from your position, which is the closest range ring from your position. Since your task is "aircraft identification," you are responsible for, and scored on, all targets within the FEBA, even though your engagement capability extends only to the 2nd range ring. Figure A-1 shows the symbols for different types of targets. A black "U" represents an "unknown;" a circle represents a "friend;" a diamond represents a "hostile." A "hexagon outline" represents a target that has been engaged by the system. Jammers are indicated by the symbol shown in Figure A-1.

You will receive 5 points for each correctly identified target, either friend or hostile. You will receive no points if you identify the target incorrectly or if you leave the target as UNKNOWN. Your goal is to maximize your point total. To help you, you will be given feedback on how well you are doing every minute. This feedback is presented in a box in the lower left-hand portion of the screen. During an attack phase, it will tell you what proportion of friends and hostiles you identified correctly, incorrectly, or left as unknown when those aircraft left the screen. If a high proportion of friendly aircraft is identified incorrectly, that means you are identifying too many friends as hostiles. Consequently, you should examine aircraft on the screen that are identified as hostile more carefully because some of them may be friends. In contrast, if a high proportion of hostile aircraft are identified incorrectly, you should examine aircraft on the screen that are identified as friends because some of them are probably hostiles.

At the end of an attack phase, the feedback box will add the proportion of friends and hostiles <u>within the FEBA</u> you identified correctly, incorrectly, or left unknown to help you determine how well you did for that phase. Attack phases vary in the amount of time they take. So, make sure to correctly identify as many friends and hostiles within the FEBA as you can throughout each phase. Remember, you get 5 points for each aircraft you correctly identify, and no points for targets you incorrectly identify or leave as "unknown."

Once you "hook" a target, its track # will appear at the top of the radar scope, as shown in Figure A-2. In order to identify a target or obtain more information about it, you must first "hook" the target by (1) using the mouse to guide a cursor to the target, and (2) pressing the mouse's left-hand button. Only one target can be hooked at a time; the target number appears at the top of the right-hand half of the screen. "Hooked targets" are represented by a square on the radar screen. <u>If you are trying to hook a target that is very close to one or more targets, keep the mouse button down until the one you want appears in the square</u>. You can then make your identification of the hooked target by using the mouse to guide the cursor to one of the three "ID" buttons in the upper portion of the left-hand half of the screen, as shown in Figure A-2. If a hooked

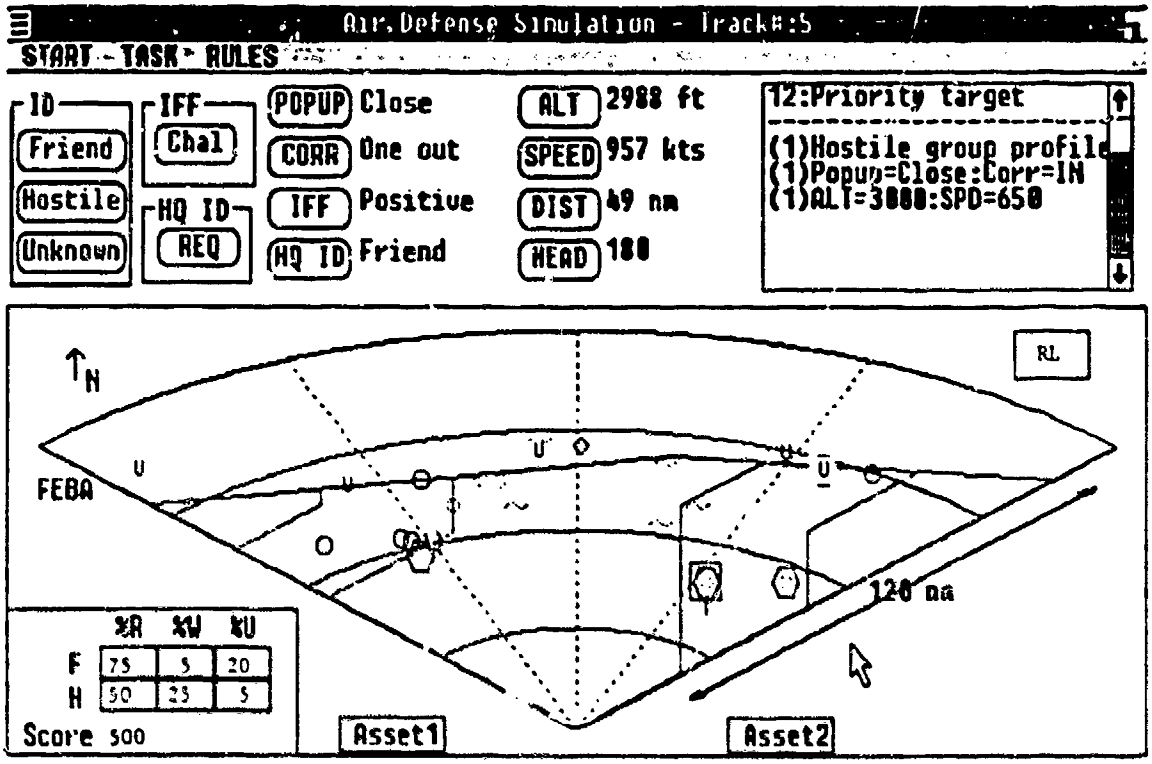


Figure A-2. The entire display screen.

target leaves the screen before you have made your identification, the track # will be replaced by "No Target" at the top of the display; simply proceed by hooking another target.

You can use the other "buttons" on the top-half of the screen to obtain more information. In particular, you can find out whether the target "popped-up" (POPUP button); whether it's "in" or "out" of the corridor (CORR button); or its speed (SPEED button), distance (DIST button), altitude (ALT button), and heading (HEAD button).

There are two buttons labeled "IFF CHAL" and "HQ ID REQ" in the upper portion of the main display. You can gather *new* information about an aircraft by pressing these buttons. "IFF CHAL" stands for "IFF Challenge," which is sending an electronic interrogation signal to which friendly aircraft can respond automatically unless their equipment has malfunctioned, they do not have an IFF transponder, or their codes are set improperly. In typical air defense systems, multiple targets can be challenged simultaneously. In the system you will be using today, an IFF challenge can be placed only against the "hooked" target. The "HQ ID REQ" button is used for you to contact higher headquarters (HQ) to ask them for the identification of the hooked target. The answer to your request will appear to the right of the button. If you receive "UNKNOWN" in response, this means that headquarters does not know the identification of the aircraft; consequently, you'll have to identify it on the basis of other information. Figure A-2 shows you all the information for the "hooked aircraft" in the right-hand safe-passage corridor.

You should give serious consideration to other available information about a target before using the "IFF CHAL" and "HQ ID REQ" buttons. Performing an IFF challenge in the "real world" opens you to exploitation (i.e., attack) by enemy aircraft. To represent this situation, you will be exploited 10% of the times you issue an IFF challenge. If you are exploited, you will lose 10 points; you will be notified in the lower left-hand corner of the display. Requesting an "HQ ID" takes time to perform in an actual air defense environment. To represent this, it will take 4 seconds for you to get a response to your request. During this time you can get information about the hooked target, but you can not hook another target. If you perform an "IFF CHAL" or request an HQ ID for a hooked target, you can recall the information from your database at a later time without incurring a point or time penalty. This is done by rehooking the target and hitting the "IFF" and "HQ" buttons located beneath the "SPEED" and "CORR" buttons.

Two Other Points. First, you will sometimes receive important information from headquarters about certain targets. This information will appear in the MESSAGE box. For example, a message might look as follows:

(1) Hostile Group Profile
(1) Popup=No:Corr=Two out
(1) Alt=11,000:Speed=1200

The number in the parentheses indicates that this is Message #1. "Hostile Group Profile" means that the message describes a group of targets that is

known to be hostile. "Popup=No:Corr=Two out" tells you that the hostiles did not pop-up and that, although they can be in both corridors, their speed *and* altitude do *not* match the parameters of the safe-passage corridors. In particular, the hostiles have an altitude of approximately 11,000 feet and a speed of approximately 1200 knots. No friendly aircraft are at that altitude and speed specified in the messages. The message is in effect for no more than ten minutes, depending upon how long it takes the hostile group to leave the radar screen. (Note: After those ten minutes, friendly aircraft may appear at the altitudes and speeds identified in the message.)

You may want to look at your message box frequently, particularly *before* hitting the "IFF CHAL" and "HQ ID" buttons to help you identify targets Remember, there is a point (i.e., "exploitation") penalty for performing an IFF challenge and a time penalty for requesting an HQ ID. You can click on the up-arrow and down-arrow to scroll up and down the message box (see Figure A-2).

Second, in a typical air defense system, you would have to acknowledge orders from higher headquarters. We have represented these orders by a light in the upper right-hand corner of the radar display. It is labeled "RL" in Figures A-1 and A-2. This light will go on throughout the session. It will stay on for 3 seconds. You must acknowledge it as fast as possible; if you fail to respond within 3 seconds, you will lose 2 points. When the light is red, you must acknowledge by pressing the middle button on the mouse. When the light is green, you must acknowledge by pressing the right-hand button on the mouse. The buttons are color-coded. You will obtain 1 point every time you respond correctly within the time limit. You will lose 1 point if you respond incorrectly, and 2 points if you fail to respond in time.

Your total "running" score will appear at the bottom of the feedback box in the lower left-hand corner of the display. The total score is a function of (1) the number of aircraft you identify correctly, (2) the number of points you lose through the exploitation of your IFF challenge, and (3) the number of points you gain or lose through your acknowledgment of the response light. The *worst possible* score you could obtain at the end of an attack phase is -400 points; the *best possible* score is +1500 points.

A number of cues are available to help you distinguish between friendly and hostile targets. Some of the cues are better discriminators than are others. The table below tells which cues are the strongest indicators and which have a lot of uncertainty associated with them. This table applies under normal circumstances--you may receive messages from time to time which indicate situations where the values do not apply. Remember that there are in general more hostile than friendly targets.

Cues Indicating FRIEND		Cues Indicating FOE		Non-Definitive
Value	Strength	Value	Strength	
IFF Positive	Very Strong	Corridor OUT	Strong	HQ ID Unknown
HQ ID Friend	Very Strong	HQ ID Hostile	Strong	Popup (all values)
Corridor IN	Strong	Jammer	Strong	Non-Jammer
Corridor One Out	Moderate	IFF No Response	Moderate	Corridor Two Out

A.3 Final Instructions

Please reread these instructions if anything is not clear and ask any questions you have before beginning. There will be practice sessions before we begin the actual session. In order to do well, you will need to pay attention to anything and everything that might help you make decisions. Please do as well as you can. The results of your work will remain anonymous and have no purpose other than this experiment.

APPENDIX B

THE "SYSTEM CAPABILITY" DESCRIPTIONS FOR THE MANUAL, OVERRIDE, AND SCREENING INTERFACES IN EXPERIMENT 1

B.1 Manual

This system keeps track of all the information about all the targets, but you must perform the identification task. When performing the identification task, remember that some aircraft will be easier to identify than others. Some aircraft will have conflicting information. As an example, an aircraft might be (a) jamming, which suggests it's hostile, but (b) giving a positive IFF, which suggests it's friendly. You can review the information about an aircraft by clicking on the appropriate buttons. In addition, you can perform an "IFF CHAL" or "HQ ID REQ" subject, of course, to the point and time penalties described in the previous section. The system will not make any ID decisions. It is up to you to consider the available information and ID each target.

B.2 Override

This system keeps track of all the information about all the targets, and helps you perform all the actions described above. The system also makes an initial identification of *all* aircraft based on (a) whether (and where) it popped-up, (b) whether it's in the corridor or not, (c) whether its speed and altitude meet the corridor parameters if it's in the corridor, and (d) whether or not it's a jammer. Aircraft initially identified as a friend are represented as *black circles*. Aircraft initially identified as hostile are represented as *black diamonds*. All jammers are initially identified as hostile. Note that the system does not have access to messages from headquarters on an HQ ID, and it can not initiate IFF challenges.

Sometimes an aircraft has conflicting information. For example, an aircraft might be (a) jamming, which suggests it's hostile, but (b) giving a positive IFF, which suggests it's friendly. Therefore, certain aircraft are easier to identify than others. The system will do the best it can for each initial identification. You can review the information the model used to identify an aircraft by hitting the appropriate buttons. In addition, you can perform an "IFF CHAL" or an "HQ ID REQ" subject, of course, to point and time penalties, respectively, described previously. The results of an IFF challenge will go directly into the system, and the system will use the results to review (and perhaps change) its identification. However, the system can *not* make direct use of the information contained in messages from headquarters, or to the response of an HQ ID. Consequently, it is quite possible for you to improve on the computer's ID results since you have access to information that the computer does not.

Unless you change the system's identification, it will represent your identification when the aircraft goes off the screen or, if the aircraft is

within the FEBA, when the attack session ends. If you make changes, they will be color-coded. In particular, a *blue circle* will represent an aircraft you identified as friend; a *red diamond* will represent an aircraft that you identified as hostile. If you change an identification to "unknown," because you want to identify another aircraft before deciding, it will be represented as a *green U*. Remember, however, that you do not get any points for aircraft identified as "unknown."

Finally, in certain cases where aircraft move into the corridor or out of the corridor, the machine might change an identification you made from hostile to friend, or vice versa. This could be either a good or bad action. It could be a good action if, for example, you made an identification before an aircraft entered a safe-passage corridor and, after the aircraft entered the corridor, the machine took this information into account. It could be a bad decision if the machine changed an identification you made on the basis of an HQ ID. Remember, the machine does not have access to the results of an HQ ID. If you identified a target as a friend on the basis of an HQ ID and the aircraft left the corridor for whatever reason, the machine could incorrectly change your identification. Whenever the machine changes an ID you made, a message will appear in the message box informing you of this change. You can 'click' on this message with the mouse to hook this target. Consequently, pay attention to "ID changes" on the few occasions they occur.

B.3 Screening

This system keeps track of all the information about all the targets, and helps you perform all the actions described above. The system also makes an initial identification of *all* aircraft based on (a) whether (and where) it popped-up, (b) whether it's in the corridor or not, (c) whether its speed and altitude meet the corridor parameters if it's in the corridor, and (d) whether or not it's a jammer. Note that the system does not have access to messages from headquarters or an HQ ID, and it can not initiate IFF challenges. On the basis of information it does have, the system uses a *blue circle* to identify aircraft that clearly appear to be friends; it uses a *red diamond* to identify aircraft that clearly appear to be hostile. If the system is less certain of its identification, it uses a *black circle* to identify "questionable friends" and a *black diamond* to identify "questionable hostiles." By "questionables" we mean there is not enough information to firmly ID, but the evidence is more in favor of one type or another. The color *black* means there is no conflicting evidence.

Sometimes an aircraft has conflicting information. For example, an aircraft might (a) be jamming, which suggests it's hostile, but (b) giving a positive IFF, which suggests it's friendly. The system uses the color "purple" to identify aircraft with conflicting information. A *purple circle* represents an aircraft that is a "questionable friend" because of conflicting information. A *purple diamond* represents an aircraft that is a "questionable hostile" because of conflicting information.

If, either because there is not enough information or the information is conflicting, the system is unable to identify the aircraft on the basis

of its initial information, the system will classify the aircraft as an "unknown" (a *black U*). Often, however, the system will indicate a "highest priority unknown." This is the target that, in the system's opinion, is the most important to ID next. The priority rating is based on the amount of uncertainty, the amount of conflict, and the aircraft's "time to nearest friendly asset." This unknown (U) will have a purple, solid circle around it. In addition, its identification number will appear at the top of the message box. You can hook the "highest priority unknown" by either (1) clicking on its identification number in the message box, which hooks it automatically, or (2) hooking it just like any other aircraft.

The system will do the best it can for each initial identification. And, unless you change the ID provided by the system, you will be scored on the basis of the ID made by the system when the aircraft goes off the screen or, if the aircraft is within the FEBA, when the attack session ends. You can review the information the model used to identify an aircraft by hitting the appropriate buttons. In addition, you can perform an "IFF CHAL" or an "HQ ID REQ" subject, of course, to the point and time penalties described previously. The results of an IFF challenge will go directly into the system and it will use it to review (and perhaps change) its identification. However, the system can *not* make direct use of the information contained in messages from headquarters, or to the response of an HQ ID. Consequently, it is quite possible for you to improve on the computer's ID results since you have access to information that the computer does not.

If you change an identification to "unknown" because you want to identify another aircraft before deciding, it will be represented as a *green U*. Remember, however, that you do not get any points for aircraft identified as "unknown."

Finally, in certain cases where aircraft move into the corridor or out of the corridor, the machine might change an identification you made from hostile to friend, or vice versa. This could be either a good or bad action. It could be a good action if, for example, you made an identification before an aircraft entered a safe-passage corridor and, after the aircraft entered the corridor, the machine took this information into account. It could be a bad decision if the machine changed an identification you made on the basis of an HQ ID. Remember, the machine does not have access to the results of an HQ ID. If you identified a target as a friend on the basis of an HQ ID and the aircraft left the corridor for whatever reason, the machine could incorrectly change your identification. Whenever the machine changes an ID you made, a message will appear in the message box informing you of this change. You can 'click' on this message with the mouse to hook this target. Consequently, pay attention to "ID changes" on the few occasions they occur.

APPENDIX C

THE SUBJECTIVE QUESTIONNAIRES USED IN BOTH EXPERIMENTS

EXPERIMENT 1

QUESTIONNAIRE

Now that you have finished all the sessions, we would like your opinion of the three systems you worked with today:

- Manual (M) (all targets appeared initially as unknowns)
- Override (O) (computer initially ID'd) all targets as friend or foe)
- Screening (S) (computer indicated conflict on some targets)

Using the scales below, please answer the questions for each of the three systems.

1. How well do you think you performed the aircraft identification task with each system? (Place the appropriate number from the scale next to each system.)

Manual _______
Override _______
Screening _______

1	2	3	4	5	6	7	8	9
Very Very Badly		Badly		Neutral		Well		Very Very Well

2. How hard did you have to work to perform the aircraft identification task with each system?

Manual _______
Override _______
Screening _______

1	2	3	4	5	6	7	8	9
Very Very Hard		Hard		Neutral		Easy		Very Very Easy

3. How much did you like performing the aircraft identification task with each system?

Manual _______
Override _______
Screening _______

1	2	3	4	5	6	7	8	9
Disliked Alot		Disliked		Neutral		Liked		Liked Alot

4. Please indicate how each cue listed below affected your ability to get correct ID's. Circle the scale point for each cue.

	Extremely Negative		Negative		Neutral		Positive		Extremely Positive
HQ ID	1	2	3	4	5	6	7	8	9
IFF	1	2	3	4	5	6	7	8	9
POP-UP	1	2	3	4	5	6	7	8	9
CORRIDOR	1	2	3	4	5	6	7	8	9
SPEED	1	2	3	4	5	6	7	8	9
ALTITUDE	1	2	3	4	5	6	7	8	9
DISTANCE	1	2	3	4	5	6	7	8	9
HEADING	1	2	3	4	5	6	7	8	9
JAMMER	1	2	3	4	5	6	7	8	9
MESSAGES	1	2	3	4	5	6	7	8	9

QUESTIONNAIRE

1. How well do you think you performed the aircraft identification task working with the system in this phase of the session?

1	2	3	4	5	6	7	8	9
Very Very Badly		Badly		Neutral		Well		Very Very Well

2. How hard did you have to work to perform the aircraft identification task working with the system in this phase of the session?

1	2	3	4	5	6	7	8	9
Very Very Hard		Hard		Neutral		Easy		Very Very Easy

3. How much did you like performing the aircraft identification task working with the system in this phase of the session?

1	2	3	4	5	6	7	8	9
Disliked Alot		Disliked		Neutral		Liked		Liked Alot

4. Comments:

EXPERIMENT 2

QUESTIONNAIRE

Now that you have finished all the sessions, we would like your opinion of the five systems you worked with today:

- Manual with Allocation (MA) (all targets appeared as unknowns unless they were identified on the basis of one of your rules)
- Override (O) (computer initially ID'd) all targets as friend or foe)
- Override with Allocation (OA) (all targets that were not identified on the basis of your rules were ID's by the computer as friend or foe)
- Screening (S) (computer indicated conflict on some targets)
- Screening with Allocation (SA) (computer indicated questionable ID's on some targets that were not ID's on the basis of your rules)

Using the scales below, please answer the questions for each of the five systems.

1. How well do you think you performed the aircraft identification task with each system? (Place the appropriate number from the scale next to each system.)

Manual with Allocation	_______
Override	_______
Override with Allocation	_______
Screening	_______
Screening with Allocation	_______

1	2	3	4	5	6	7	8	9
Very Very Badly		Badly		Neutral		Well		Very Very Well

2. How hard did you have to work to perform the aircraft identification task with each system?

Manual with Allocation	_______
Override	_______
Override with Allocation	_______
Screening	_______
Screening with Allocation	_______

1	2	3	4	5	6	7	8	9
Very Very Hard		Hard		Neutral		Easy		Very Very Easy

3 How much did you like performing the aircraft identification task with each system?

Manual with Allocation	_______
Override	_______
Override with Allocation	_______
Screening	_______
Screening with Allocation	_______

1	2	3	4	5	6	7	8	9
Disliked Alot		Disliked		Neutral		Liked		Liked Alot

4. How did the ability to establish your own rules affect your aircraft identification performance with each system?

Manual with Allocation	_______
Override with Allocation	_______
Screening with Allocation	_______

1	2	3	4	5	6	7	8	9
Extremely Negative		Negative		Neutral		Positive		Extremely Positive

5. Please indicate how each cue listed below affected your ability to get correct ID's. Circle the scale point for each cue.

	Extremely Negative		Negative		Neutral		Positive		Extremely Positive
HQ ID	1	2	3	4	5	6	7	8	9
IFF	1	2	3	4	5	6	7	8	9
POP-UP	1	2	3	4	5	6	7	8	9
CORRIDOR	1	2	3	4	5	6	7	8	9
SPEED	1	2	3	4	5	6	7	8	9
ALTITUDE	1	2	3	4	5	6	7	8	9
DISTANCE	1	2	3	4	5	6	7	8	9
HEADING	1	2	3	4	5	6	7	8	9
JAMMER	1	2	3	4	5	6	7	8	9
MESSAGES	1	2	3	4	5	6	7	8	9

APPENDIX D

THE "SYSTEM CAPABILITY" DESCRIPTIONS FOR THE THREE ALLOCATION CONDITIONS IN EXPERIMENT 2

D.1 Manual with Allocation

This system keeps track of all the information about all the targets, but you must perform the identification task unless you direct the system to do otherwise. Initially all targets are classified as "unknown;" it is your job to classify every target as either "friend" or "hostile." When performing the identification task, remember that some aircraft will be easier to identify than others. Some aircraft will have conflicting information. As an example, an aircraft might be (a) jamming, which suggests it's hostile, but (b) giving a positive IFF, which suggests it's friendly. You can review the information about an aircraft by clicking on the appropriate buttons. In addition, you can perform an "IFF CHAL" or HQ ID REQ" subject, of course, to the point and time penalties described in the previous section. It is up to you to consider the available information and ID each target.

In this system, you can also create "identification rules" to help you by allowing the system to ID targets meeting specified criteria. For example, you can tell the system to identify all jammers as hostile, and the system will do so automatically. Or, for example, you could tell the system to automatically identify all aircraft in a safe-passage corridor with the correct speed and altitude as friends, and so forth.

Before beginning an attack phase, you will have the opportunity to create identification rules. The system will start off with the display shown in Figure D-1. Table D-1 identifies the options for each button in the rule-creation component of the system. For example, the POPUP button has four options: No, Close, Feba, and Yes. The CORR button has five options: In, One-Out, Two-Out, Out, and N/A. Let's assume, for example, you wanted to say that all jammers are to be identified as hostile. You would move the mouse to the JAMM button and click the left mouse button. When YES comes up, indicating that you are referring to jammers, you would then go over to the RESULT column in the far left-hand corner of the display and click-on "Hostile," implying that you want all jammers to be identified as hostile. Then, click-on "Save Rule," which is directly below the "Hostile" button to save your identification rule. The system now understands that all jammers are to be identified automatically as hostile.

If you now click-on RULES at the top of the display, making sure to hold down the mouse button, you will see that a rule called R6(H) has been created. At any time you may move the mouse over R6(H) and lift your finger off the mouse button. The values of the R6(H) identification rule will then appear on the screen; that is, "jammers are hostile." If you want to erase this rule, just click-on "Clear Rule" directly below "Save Rule." As another example, if you wanted all aircraft doing everything

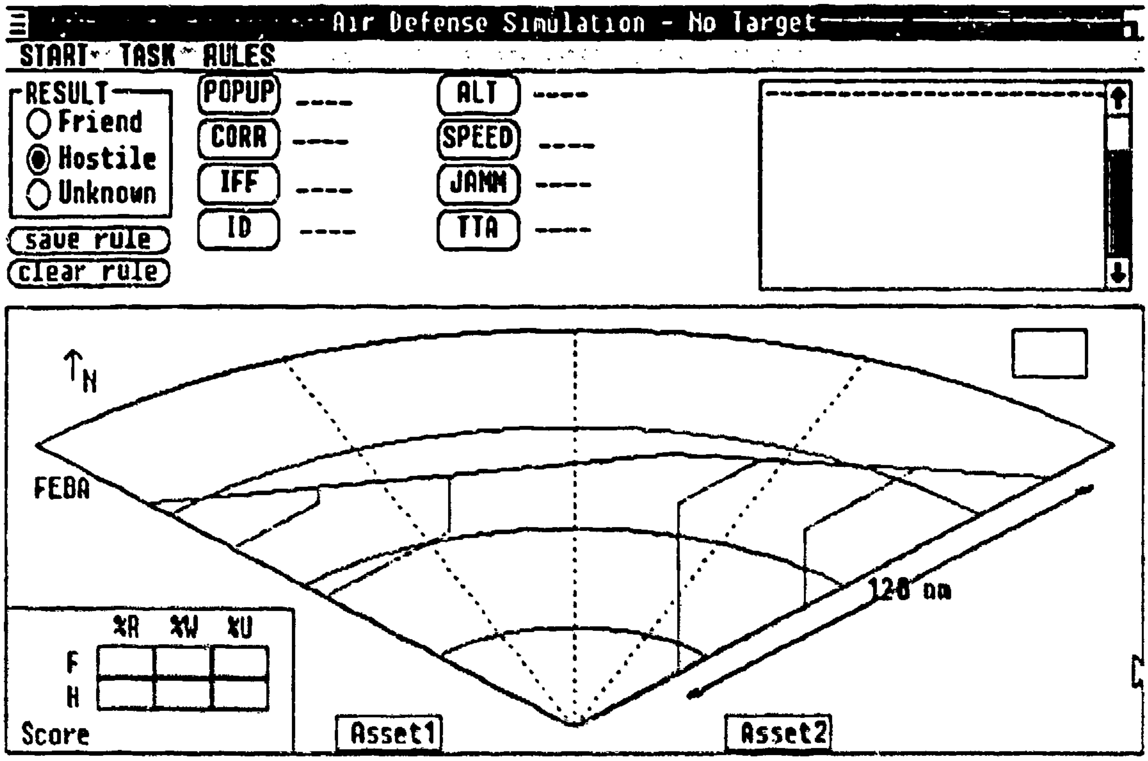


Figure D-1. Initial screen display.

Table D-1

Rule Work Sheet

Rule # ___

ID as:	FRIEND	
	HOSTILE	
	UNKNOWN	

Criteria

POPUP :	NO	(Target did not pop up)
	CLOSE	(Target popped up close to friendly asset)
	FEBA	(Target popped up near FEBA)
	YES	(Target popped up close *or* near FEBA)
CORR :	IN	(Target in corridor, speed *and* altitude are correct)
	ONE OUT	(Target in corridor, speed *or* altitude incorrect)
	TWO OUT	(Target in corridor, speed *and* altitude incorrect)
	OUT	(Target not in corridor)
	N/A	(Targets have not reached corridor entrance)
IFF :	POSITIVE	(Targets that *have* been challenged, respond as FRIEND)
	NEGATIVE	(Targets that *have* been challenged, no response)
	NO CHAL	(Targets that have not been challenged)
ID :	UNKNOWN	(All unknown targets)
ALT : (altitude band, feet)		0:1000
		1000:2000
		2000:5000
		5000:10000
		10000:60000
		60000+
SPEED : (knots)		0:200
		200:400
		400:600
		600:800
		800:1000
		1000+
JAMM :	YES	(All targets that are jamming)
	NO	(All targets that are not jamming)
TTA . (Time to Asset)	1 MIN	(All targets within 1 minute from friendly assets)
	2 MIN	(All targets within 2 minutes from friendly assets)
	3 MIN	(All targets within 3 minutes from friendly assets)
	3 MIN+	(All targets more than 3 minutes from friendly assets)

correct in the corridor to be identified as friend, you would do the following:

- click-on CORR until it reads In;
- click-on Friend; and
- click-on Save Rule.

Before creating a new rule, you need to select a rule number. To do this, (1) click-on RULES, and (2) while pressing the mouse button, move the mouse to an empty rule. Now release the mouse button. Initially rule R6 is selected for you. The first five spaces for saving rules in the RULES box are listed under MESSAGE to indicate that these rules will be executed first by the system because they refer to rules you created to identify a group of aircraft with the characteristics indicated in a message. For example, assume you received the message below:

(1) Hostile Group Profile
(1) Popup=No: Corr=TwoOut
(1) Alt=11,000: Speed=1200

You could create a rule that identified all these aircraft as hostile by doing the following:

- select Rule MR1;
- click-on POPUP until it reads No;
- click-on CORR until it reads Two Out;
- click-on ALT until it reads 10000-80000;
- click-on SPEED until it reads 1000+;
- click-on Hostile; and
- click-on Save Rule.

This rule will be stored as MR1(H) in the Message category under the RULES menu. It will identify all aircraft as hostile with the above characteristics. You should erase or 'clear' this rule from the RULES menu when this message (i.e., Message 1) disappears from the Message box because all the hostile aircraft with these (message) characteristics have left the screen. Note that a slot exists for up to five messages; place each message in the slot with the same number. For example, a rule for message #3 would go in the slot labeled MR3.

You can create as many rules as you like before and during an attack phase. The message rules always get executed first. The other rules get executed in the order that they are listed. For example, MR1 (a message rule) gets executed before R6, which gets executed before R10. <u>This order is important</u>. If you know that you want all jammers to be hostile and any other aircraft that are traveling correctly within a corridor to be friendly, then the rule for jammers should come before the corridor rule.

You can use the system's rule-creation component at any time. If you think you would like to have certain rules saved in the system prior to beginning the attack phase, save them now. After you have executed these rules and are ready to begin, click-on START at the top of the display to begin the attack phase. This will automatically bring up the standard dis-

play to be used for making target identifications. The attack phase can not be stopped once it is started. You can select the RULES option at any time during the session.

Any target not covered by a rule is identified as an unknown.. You will be scored on the basis of the ID made by you or by the system using your rules when the aircraft goes off the screen (or, if the aircraft is within the FEBA, when the attack session ends). You can review the information the system used to identify an aircraft by hitting the appropriate buttons. In addition, you can perform an "IFF CHAL" or an "HQ ID REQ" subject, of course, to the point and time penalties described previously. The results of an IFF challenge will go directly into the system and it will use it to review (and perhaps change) its identification. However, the system can *not* make direct use of the information contained in messages from headquarters, or to the response of an HQ ID. Consequently, it is quite possible for you to improve on the computer's ID results since you have access to information that the computer does not.

If you ID a target, it will be color-coded. In particular, a *blue circle* will represent an aircraft you identified as friend; a *red diamond* will represent an aircraft that you identified as hostile. If you change an identification to "unknown," because you want to identify another aircraft before deciding, it will be represented as a *green U*. Note that you will receive no points for any target left unknown.

D.2 Override with Allocation

This system keeps track of all the information about all the targets, and helps you perform all the actions described above. The system also makes an initial identification of *all* aircraft based on (a) whether (and where) it popped-up, (b) whether it's in the corridor or not, (c) whether its speed and altitude meet the corridor parameters if it's in the corridor, and (d) whether or not it's a jammer. Aircraft initially identified as a friend are represented as *black circles*. Aircraft initially identified as hostile are represented as *black diamonds*. All jammers are initially identified as hostile. Note that the system does not have access to messages from headquarters on an HQ ID, and it can not initiate IFF challenges.

Sometimes an aircraft has conflicting information. For example, an aircraft might be (a) jamming, which suggests it's hostile, but (b) giving a positive IFF, which suggests it's friendly. Therefore, certain aircraft are easier to identify than others. The system will do the best it can for each initial identification. You can review the information the model used to identify an aircraft by hitting the appropriate buttons. In addition, you can perform an "IFF CHAL" or an "HQ ID REQ" subject, of course, to point and time penalties, as described previously. The results of an IFF challenge will go directly into the system, and the system will use the results to review (and perhaps change) its identification. However, the system can *not* make direct use of the information contained in messages from headquarters, or to the response of an HQ ID. Consequently, it is

quite possible for you to improve on the computer's ID results since you have access to information that the computer does not.

In this system, you can also create "identification rules" to help you by allowing the system to ID targets meeting specified criteria. For example, you can tell the system to identify all jammers as hostile, and the system will do so automatically. Or, for example, you could tell the system to automatically identify all aircraft in a safe-passage corridor with the correct speed and altitude as friends, and so forth.

Before beginning an attack phase, you will have the opportunity to create identification rules. The system will start off with the display shown in Figure D-2. Table D-2 identifies the options for each button in the rule-creation component of the system. For example, the POPUP button has four options: No, Close, Feba, and Yes. The CORR button has five options: In, One-Out, Two-Out, Out, and N/A. Let's assume, for example, you wanted to say that all jammers are to be identified as hostile. You would move the mouse to the JAMM button and click the left mouse button. When YES comes up, indicating that you are referring to jammers, you would then go over to the RESULT column in the far left-hand corner of the display and click-on "Hostile," implying that you want all jammers to be identified as hostile. Then, click-on "Save Rule," which is directly below the "Hostile" button to save your identification rule. The system now understands that all jammers are to be identified automatically as hostile.

If you now click-on RULES at the top of the display, making sure to hold down the mouse button, you will see that a rule called R6(H) has been created. At any time you may move the mouse over R6(H) and lift your finger off the mouse button. The values of the R6(H) identification rule will then appear on the screen; that is, "jammers are hostile." If you want to erase this rule, just click-on "Clear Rule" directly below "Save Rule." As another example, if you wanted all aircraft doing everything correct in the corridor to be identified as friend, you would do the following:

- click-on CORR until it reads In;
- click-on Friend; and
- click-on Save Rule.

Before creating a new rule, you need to select a rule number. To do this, (1) click-on RULES, and (2) while pressing the mouse button, move the mouse to an empty rule. Now release the mouse button. Initially rule R6 is selected for you. The first five spaces for saving rules in the RULES box are listed under MESSAGE to indicate that these rules will be executed first by the system because they refer to rules you created to identify a group of aircraft with the characteristics indicated in a message. For example, assume you received the message below:

(1) Hostile Group Profile
(1) Popup=No: Corr=TwoOut
(1) Alt=11,000· Speed=1200

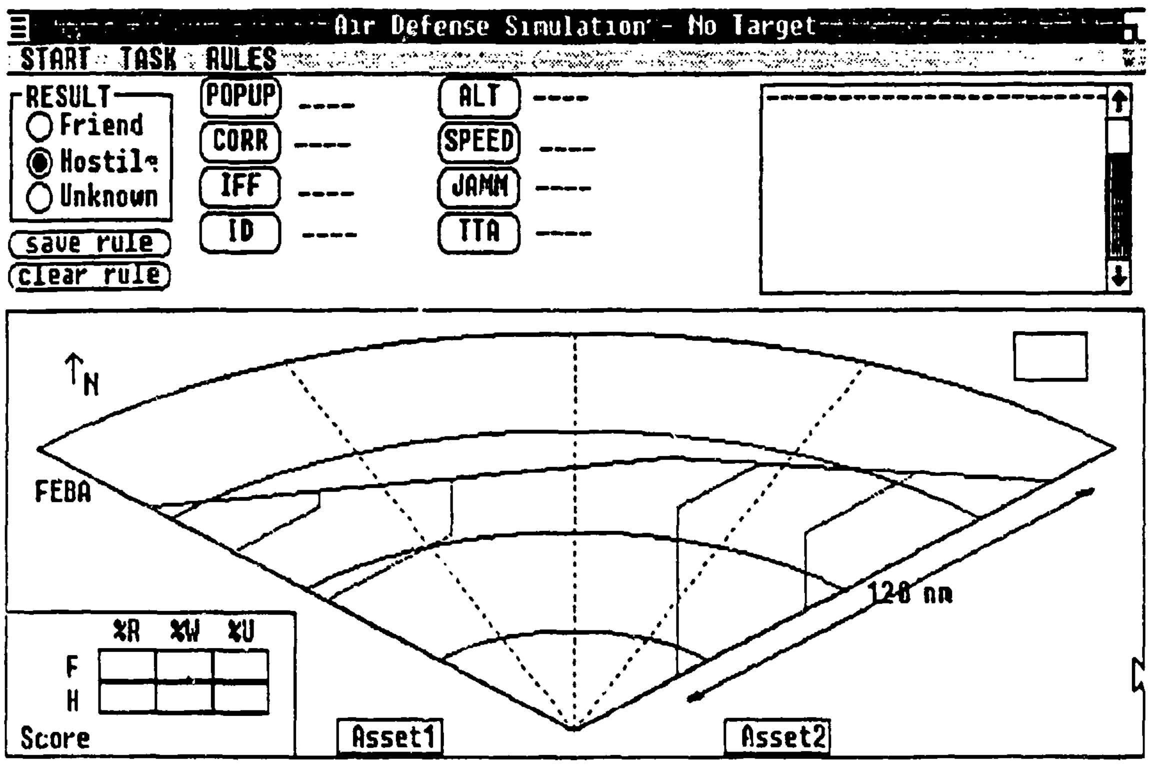


Figure D-2. Initial screen display.

Table D-2

Rule Work Sheet

Rule # ___

ID as:	FRIEND HOSTILE UNKNOWN	

Criteria

POPUP :	NO	(Target did not pop up)
	CLOSE	(Target popped up close to friendly asset)
	FEBA	(Target popped up near FEBA)
	YES	(Target popped up close *or* near FEBA)
CORR :	IN	(Target in corridor, speed *and* altitude are correct)
	ONE OUT	(Target in corridor, speed *or* altitude incorrect)
	TWO OUT	(Target in corridor, speed *and* altitude incorrect)
	OUT	(Target not in corridor)
	N/A	(Targets have not reached corridor entrance)
IFF :	POSITIVE	(Targets that *have* been challenged, respond as FRIEND)
	NEGATIVE	(Targets that *have* been challenged, no response)
	NO CHAL	(Targets that have not been challenged)
ID:	UNKNOWN	(All unknown targets)
ALT: (altitude band, feet)		0:1000 1000:2000 2000:5000 5000:10000 10000:80000 80000+
SPEED : (knots)		0:200 200:400 400:600 600:800 800:1000 1000+
JAMM :	YES	(All targets that are jamming)
	NO	(All targets that are not jamming)
TTA : (Time to Asset)	1 MIN	(All targets within 1 minute from friendly assets)
	2 MIN	(All targets within 2 minutes from friendly assets)
	3 MIN	(All targets within 3 minutes from friendly assets)
	3 MIN+	(All targets more than 3 minutes from friendly assets)

You could create a rule that identified all these aircraft as hostile by doing the following:

- select Rule MR1;
- click-on POPUP until it reads No;
- click-on CORR until it reads Two Out;
- click-on ALT until it reads 10000-80000;
- click-on SPEED until it reads 1000+;
- click-on Hostile; and
- click-on Save Rule.

This rule will be stored as MR1(H) in the Message category under the RULES menu. It will identify all aircraft as hostile with the above characteristics. You should erase or 'clear' this rule from the RULES menu when this message (i.e., Message 1) disappears from the Message box because all the hostile aircraft with these (message) characteristics have left the screen. Note that a slot exists for up to five messages; place each message in the slot with the same number. For example, a rule for message #3 would go in the slot labeled MR3.

You can create as many rules as you like before and during an attack phase. The message rules always get executed first. The other rules get executed in the order that they are listed. For example, MR1 (a message rule) gets executed before R6, which gets executed before R10. This order is important. If you know that you want all jammers to be hostile and any other aircraft that are traveling correctly within a corridor to be friendly, then the rule for jammers should come before the corridor rule.

You can use the system's rule-creation component at any time. If you think you would like to have certain rules saved in the system prior to beginning the attack phase, save them now. After you have executed these rules and are ready to begin, click-on START at the top of the display to begin the attack phase. This will automatically bring up the standard display to be used for making target identifications. The attack phase can not be stopped once it is started. You can select the RULES option at any time during the session.

Unless you change the system's identification, it will represent your identification when the aircraft goes off the screen or, if the aircraft is within the FEBA, when the attack session ends. If you make changes, they will be color-coded. In particular, a *blue circle* will represent an aircraft you identified as friend; a *red diamond* will represent an aircraft that you identified as hostile. If you change an identification to "unknown," because you want to identify another aircraft before deciding, it will be represented as a *green U*. The system ensures that all "unknowns" are identified before they leave the screen. If you do not have an opportunity to identify an aircraft, or have identified an aircraft as "unknown," the system will make an identification on the basis of whatever information is available about that aircraft.

D.3 Screening with Allocation

This system keeps track of all the information about all the targets, and helps you perform all the actions described above. The system also makes an initial identification of *all* aircraft based on (a) whether (and where) it popped-up, (b) whether it's in the corridor or not, (c) whether its speed and altitude meet the corridor parameters if it's in the corridor, and (d) whether or not it's a jammer. Note that the system does not have access to messages from headquarters or an HQ ID, and it can not initiate IFF challenges. On the basis of information it does have, the system uses a *blue circle* to identify aircraft that clearly appear to be friends; it uses a *red diamond* to identify aircraft that clearly appear to be hostile.

If the system is less certain of its identification, it uses a *black circle* to identify "questionable friends" and a *black diamond* to identify "questionable hostiles." By "questionables" we mean there is either (a) not enough information to firmly ID or (b) the information is conflicting, but the evidence is more in favor of one ID or another. For example, an aircraft in the corridor with the wrong speed or altitude might be identified as a questionable friend.

If, either because there is not enough information or the information is conflicting, the system is unable to identify the aircraft on the basis of its initial information, the system will classify the aircraft as an "unknown" (a *black U*). Often the system will indicate a "highest priority unknown." This is the target that, in the system's opinion, is the most important to ID next. The priority rating is based on the amount of uncertainty, the amount of conflict, and the aircraft's "time to nearest friendly asset." This unknown (U) will have a solid purple circle around it. In addition, its identification number will appear at the top of the message box. You can hook the "highest priority unknown" by either (1) clicking on its identification number in the message box, which hooks it automatically, or (2) hooking it just like any other aircraft.

In this system, you can also create "identification rules" to help you by allowing the system to ID targets meeting specified criteria. For example, you can tell the system to identify all jammers as hostile, and the system will do so automatically. Or, for example, you could tell the system to automatically identify all aircraft in a safe-passage corridor with the correct speed and altitude as friends, and so forth.

Before beginning an attack phase, you will have the opportunity to create identification rules. The system will start off with the display shown in Figure D-3. Table D-3 identifies the options for each button in the rule-creation component of the system. For example, the POPUP button has four options: No, Close, Feba, and Yes. The CORR button has five options: In, One-Out, Two-Out, Out, and N/A. Let's assume, for example, you wanted to say that all jammers are to be identified as hostile. You would move the mouse to the JAMM button and click the left mouse button. When YES comes up, indicating that you are referring to jammers, you would then go over to the RESULT column in the far left-hand corner of the display and click-on "Hostile," implying that you want all jammers to be identified as

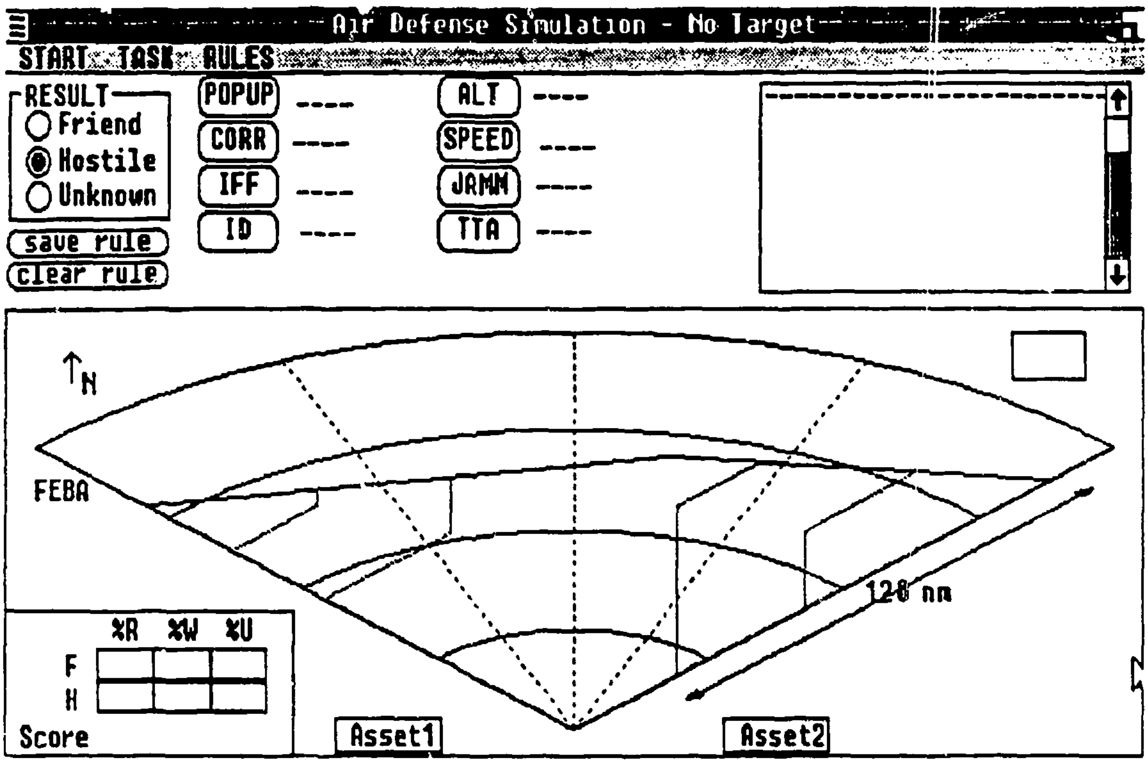


Figure D-3. Initial screen display.

Table D-3

Rule Work Sheet

Rule # ___

ID as:	FRIEND HOSTILE UNKNOWN	

Criteria

POPUP :	NO	(Target did not pop up)
	CLOSE	(Target popped up close to friendly asset)
	FEBA	(Target popped up near FEBA)
	YES	(Target popped up close *or* near FEBA)
CORR :	IN	(Target in corridor, speed *and* altitude are correct)
	ONE OUT	(Target in corridor, speed *or* altitude incorrect)
	TWO OUT	(Target in corridor, speed *and* altitude incorrect)
	OUT	(Target *not* in corridor)
	N/A	(Targets have not reached corridor entrance)
IFF :	POSITIVE	(Targets that *have* been challenged, respond as FRIEND)
	NEGATIVE	(Targets that *have* been challenged, no response)
	NO CHAL	(Targets that have not been challenged)
ID :	UNKNOWN	(All unknown targets)
	QUESTION	(All targets with a questionable ID)
	NOT FIRM	(Any target that is either unknown or questionable)
ALT : (altitude band, feet)		0:1000 1000:2000 2000:5000 5000:10000 10000:80000 80000+
SPEED : (knots)		0:200 200:400 400:600 600:800 800:1000 1000+
JAMM :	YES	(All targets that are jamming)
	NO	(All targets that are not jamming)
TTA : (Time to Asset)	1 MIN	(All targets within 1 minute from friendly assets)
	2 MIN	(All targets within 2 minutes from friendly assets)
	3 MIN	(All targets within 3 minutes from friendly assets)
	3 MIN+	(All targets more than 3 minutes from friendly assets)

hostile. Then, click-on "Save Rule," which is directly below the "Hostile" button to save your identification rule. The system now understands that all jammers are to be identified automatically as hostile.

If you now click-on RULES at the top of the display, making sure to hold down the mouse button, you will see that a rule called R6(H) has been created. At any time you may move the mouse over R6(H) and lift your finger off the mouse button. The values of the R6(H) identification rule will then appear on the screen; that is, "jammers are hostile." If you want to erase this rule, just click-on "Clear Rule" directly below "Save Rule." As another example, if you wanted all aircraft doing everything correct in the corridor to be identified as friend, you would do the following:

- click-on CORR until it reads In;
- click-on Friend; and
- click-on Save Rule.

Before creating a new rule, you need to select a rule number. To do this, (1) click-on RULES, and (2) while pressing the mouse button, move the mouse to an empty rule. Now release the mouse button. Initially rule R6 is selected for you. The first five spaces for saving rules in the RULES box are listed under MESSAGE to indicate that these rules will be executed first by the system because they refer to rules you created to identify a group of aircraft with the characteristics indicated in a message. For example, assume you received the message below:

(1) Hostile Group Profile
(1) Popup=No: Corr=TwoOut
(1) Alt=11,000: Speed=1200

You could create a rule that identified all these aircraft as hostile by doing the following:

- select Rule MR1;
- click-on POPUP until it reads No;
- click-on CORR until it reads Two Out;
- click-on ALT until it reads 10000-80000;
- click-on SPEED until it reads 1000+;
- click-on Hostile; and
- click-on Save Rule.

This rule will be stored as MR1(H) in the Message category under the RULES menu. It will identify all aircraft as hostile with the above characteristics. You should erase or 'clear' this rule from the RULES menu when this message (i.e., Message 1) disappears from the Message box because all the hostile aircraft with these (message) characteristics have left the screen. Note that a slot exists for up to five messages; place each message in the slot with the same number. For example, a rule for message #3 would go in the slot labeled MR3.

You can create as many rules as you like before and during an attack phase. The message rules always get executed first. The other rules get

executed in the order that they are listed. For example, MR1 (a message rule) gets executed before R6, which gets executed before R10. This order is important. If you know that you want all jammers to be hostile and any other aircraft that are traveling correctly within a corridor to be friendly, then the rule for jammers should come before the corridor rule.

You can use the system's rule-creation component at any time. If you think you would like to have certain rules saved in the system prior to beginning the attack phase, save them now. After you have executed these rules and are ready to begin, click-on START at the top of the display to begin the attack phase. This will automatically bring up the standard display to be used for making target identifications. The attack phase can not be stopped once it is started. You can select the RULES option to create or delete rules at any time during the session.

The system will do the best it can for each initial identification both on its own and with the rules you provide. And, unless you change the ID provided by the system, you will be scored on the basis of the ID made by the system when the aircraft goes off the screen or, if the aircraft is within the FEBA, when the attack session ends. You can review the information the model used to identify an aircraft by hitting the appropriate buttons. In addition, you can perform an "IFF CHAL" or an "HQ ID REQ" subject, of course, to the point and time penalties described previously. The results of an IFF challenge will go directly into the system and it will use it to review (and perhaps change) its identification. However, the system can *not* make direct use of the information contained in messages from headquarters, or to the response of an HQ ID. Consequently, it is quite possible for you to improve on the computer's ID results since you have access to information that the computer does not.

If you ID a target, it will be color-coded. In particular, a *blue circle* will represent an aircraft you identified as friend; a *red diamond* will represent an aircraft that you identified as hostile. If you change an identification to "unknown," because you want to identify another aircraft before deciding, it will be represented as a *green U*. The system ensures that all "unknowns" are identified before they leave the screen. If you do not have an opportunity to identify an aircraft, or have identified an aircraft as "unknown," the system will make an identification on the basis of whatever information is available about that aircraft.

APPENDIX E

CUE-PROCESSING SEQUENCE
FOLLOWED BY ONE OF THE DSC DOMAIN EXPERTS IN THE NONALLOCATION CONDITIONS
FOR THE MANUAL, OVERRIDE, AND SCREENING CONDITIONS

E.1 Manual Non-Message

1. Challenge all
   If positive response → F
   If no response
      Ask HQ
         If Friend → F
         If Hostile → H
         If Unknown → H

2. All out of corridor → H
   If in corridor (lateral boundaries)
      Challenge
         If positive → F
         If no response
            If Friend → F
            If Hostile → H
            If Unknown → H

3. All out of corridor → H
   If in corridor (lateral boundaries)
      Ask HQ
         If Friend → F
         If Hostile
            Challenge
               If positive → F
               If no response → H

4. All out of corridor → H
   If in corridor (lateral boundaries)
      If Two-Out
         Challenge
            If no response → H
            If positive → F
      If One-Out
         Ask HQ ID
            If Friend → F
            If Hostile
               Challenge

   If positive → F

   If no response → H

E.2 Override

The computer has already accounted for One-Out, Two-Out, etc. factors; therefore, I only focus on new information, e.g., IFF, HQ ID

1. If computer IDs as Friend
    - Challenge
        - If no response → H
        - If positive → no change

   If computer IDs as Foe
    - Challenge
        - If no response → no change
        - If positive → F

2. If computer IDs as Friend
    - Ask HQ ID
        - If Friend → no change
        - If Hostile
            - Challenge
                - If Friend → no change
                - If no response → H

   If computer IDs as Hostile
    - Ask HQ ID
        - If Hostile → no change
        - If Friend → F

E.3 Screening

ID on red and blue targets is pretty firm; therefore, concentrate on black targets only.

Apply strategies for override to black targets.